Univerza v Ljubljani
Fakulteta za računalništvo in informatiko

Urška Kosec

Napovedovanje čustvene naravnanosti avtorjev v spletnih komentarjih

diplomsko delo

univerzitetni študijski program računalništvo in informatika

Mentor: prof. dr. Blaž Zupan

Ljubljana 2014

Fakulteta za računalništvo in informatiko izdaja naslednjo nalogo:

Tematika naloge:

V nalogi preučite, ali je za komentarje spletnih novic moč napovedati čustveno naravnanost pisca komentarja iz zapisanega besedila. Ker so ti komentarji tipično kratki, pristop strojnega učenja zasnujte tako, da besedila predstavi z n-terkami znakov. Na izbranem praktičnem primeru preizkusite in ocenite napovedne točnosti različnih tehnik strojnega učenja. Poročajte o uspešnosti pristopa.

Izjava o avtorstvu diplomskega dela

Spodaj podpisana Urška Kosec, z vpisno številko **63070102**, sem avtorica diplomskega dela z naslovom:

Napovedovanje čustvene naravnanosti avtorjev v spletnih komentarjih

S svojim podpisom zagotavljam, da:

- sem diplomsko delo izdelala samostojno pod mentorstvom prof. dr. Blaža Zupana,
- so elektronska oblika diplomskega dela, naslov (slov., angl.), povzetek (slov., angl.) ter ključne besede (slov., angl.) identični s tiskano obliko diplomskega dela,
- soglašam z javno objavo elektronske oblike diplomskega dela na svetovnem spletu preko univerzitetnega spletnega arhiva.

V Ljubljani, dne 10. maja 2014 Podpis avtorja:

Najlepša hvala prof. dr. Blažu Zupanu za zelo dobro in korektno mentorstvo ter usmerjanje s koristnimi nasveti pri diplomskem delu. Zahvaljujem se tudi ostalim, ki so me spodbujali pri uresničitvi tega cilja.

Kazalo

Seznam uporabljenih kratic

kratica	**angleško**	**slovensko**
LR	logistic regression	logistična regresija
SVM	support vector machine	metoda podpornih vektorjev
KNN	k-nearest neighbours	k najbližjih sosedov
RF	random forests	metoda naključnih gozdov
ZP	winning parameter	zmagovalni parameter
OP	parameter score	ocena parametra
OUM	score on the training set	ocena na učni množici

Povzetek

V nalogi smo raziskali napovedljivost sentimentalnega pridiha oziroma čustvene naravnanosti avtorjev v komentarjih spletnih novic. Na področju tovrstne analize besedil je bilo v preteklih letih objavljeno večje število sorodnih raziskav za angleški jezik, a ker za slovenščino, razen v nedavni diplomski nalogi na UL FRI, podobnih raziskav nismo zasledili, je to glede na vse posebnosti slovenskega jezika za našo nalogo predstavljalo še dodatni izziv. Kratka besedila smo želeli čimbolj točno razvrstiti v kategoriji pozitivnih oziroma negativnih komentarjev. Preučili smo, kako se ta problem razlikuje od klasičnega razvrščanja besedil glede na temo in kakšne so podobnosti med problemskima domenama. V nalogi ugotovimo, da uporabljene tehnike strojnega učenje ne dosegajo pričakovanih rezultatov. Možen razlog za takšno odstopanje je predstavitev besedil z n-terkami znakov, ki ne upošteva semantike besedila oziroma besed, iz katerih je komentar sestavljen ter ne upošteva njihovih morebitnih interakcij. Dodatna težavnost pri obravnavani nalogi so tudi zelo kratki komentarji.

Ključne besede: napovedovanje čustvene naravnanosti, rudarjenje mnenj, odkrivanje znanj iz podatkov, strojno učenje, n-terka, klasifikacijske metode, logloss, ocena točnosti, logistična regresija, metoda podpornih vektorjev, metoda k najbližjih sosedov, metoda naključnih gozdov, skladanje.

Abstract

The project described in this Thesis dealt with machine learning-based classification of the sentimental impact and emotional affection of the comments posted with news articles in Slovene language on the web. In the past years sentiment analysis has become an important research topics with substantial number of publications for texts in English language, while for the Slovene language, except in the recent thesis at the University of Ljubljana, Faculty of Computer and Information science, the topic has not been explored well. In relation to all the features of the Slovenian language this represented an additional challenge. Our goal was to determine, if a machine learning algorithm can correctly classify these comments as positive or negative. We examined how this problem differs from the classical topical classification of texts and what are the similarities between problem domains. Our work shows that the problem is hard and that a typical application of machine learning based on k-mer representation of text does not yield the expected results. A possible reason for poor predictive performance may be lack of semantic information in such representation. Also, many of the texts we have included in our analysis were very short.

Keywords: sentiment prediction, opinion mining, data mining, machine learning, k-mer, classification methods, logloss, accuracy score, logistic regression, support vector machines, k-nearest neighbours, random forests, stacking.

Poglavje 1

Uvod

Skozi čas so se do danes na različnih spletnih medijih nabrale zelo velike zbirke besedilnih podatkov. Ker razvoj sodobne tehnologije stremi k digitalizaciji vseh podatkov in dostopnosti le-teh, je internet postal že prava zakladnica različnih dokumentov. Za boljšo organizacijo tako velikih podatkovnih naborov se raziskovalci, razvijalci ter ponudniki spletnih stranih trudijo, da bi uporabnikom obogatili uporabniško izkušnjo z dodatno analizo besedil ter pridobivanjem dodatnih informacij iz nestrukturiranih dokumentov.

Začetki raziskovanja in razvrščanja besedil v skupine so se začeli z napovedovanjem tem, o katerih besedila govorijo. Eno takih razvrščanj med drugim opišejo tudi Getoor in sodelavci na primeru razvrščanj spletnih strani na portalih glede na tematike, o kateri besedila govorijo [7].

V zadnjem času pa je zaradi potrebe po izražanju mnenja vsakega posameznika in zaradi enostavnosti le-tega v virtualnem svetu nastalo precej forumov pa tudi socialnih omrežij, kjer posamezniki med seboj diskutirajo in izmenjujejo svoje poglede na določeno temo. Pri tem se mnogokrat njihovi komentarji ne osredotočajo več na samo temo, ki predstavlja ozadje diskusije, marveč bolj na predhodno objavljene komentarje na forumih. Tovrstne diskusije mnogokrat postanejo čustveno nabite, pri branju komentarjev pa postane jasno, da se del diskutantov do teme ali pa do komentarjev izraža

pozitivno in se na primer z vsebino osnovnega prispevka strinja, drugi del pa morda kaže odklonilen odnos do obravnavane tematike. Odprto vprašanje je, ali in do kakšne stopnje lahko to čustveno naklonjenost komentatorjev avtomatsko razberemo iz njihovih komentarjev. Torej, ali se je možno naučiti napovednega modela, ki bi komentarje lahko na podlagi zapisanega besedila razvrstil med pozitivne in negativne. Raziskovanje na tem področju lahko pripomore pri sistemih poslovne inteligence in optimizacije poslovnih procesov ali priporočilnih sistemih, kjer bi bilo možno zajeti mnenje uporabnikov v naravnem jeziku in iz tega avtomatično izluščiti vsa potrebna dejstva.

Začetnim člankom o raziskovanju sentimenta [2, 12] na angleških besedilih je do danes sledilo že precej raziskav na tem področju, za slovenščino pa z izjemo nedavnega diplomskega dela na sorodno temo [3] še nismo zasledili konkretnih raziskav. Slovenski jezik lahko zaradi svojih specifičnosti predstavlja še dodaten izziv.

V pričujoči diplomski nalogi je bila naša naloga preučiti, kako uspešne so lahko različne metode strojnega učenja pri razvrščanju slovenskih besedil v pozitivni ali negativni razred. Naš cilj je bil gradnja napovednih modelov, ki bi iz besedila komentarja zaznala naklonjenost avtorjev komentarjev k temi članka, na katerega se komentarji nanašajo. S podobni problemom sta se za angleški jezik ukvarjala že Pang in Lee [2] ter preučevale naravnanost komentatorjev pri kritikah filmov. V tej nalogi pa smo k problemu pristopili na malo drugačen način, saj ne uporabimo prej pripravljenega korpusa subjektivnosti ali leksikona, kot so to implemntirali Mihalcea in sodelavci [13]. Prav tako se ne osredotočamo na kontekstualne fraze, ki lahko kažejo na določeno nagibanje (Turney [12]). Cilj naše naloge je bil namreč odkriti, ali lahko, za slovenski jezik, dober napovedni model tehnike strojnega učenja odkrijejo popolnoma avtomatsko, brez dodatnega semantičnega predznanja. V ta namen smo besedila predstavili atributno, s frekvencami n-terke črk. Podobna predstavitev se standardno uporablja pri klasifikaciji besedil, na primer na

področju odkrivanja nezaželene pošte [5].

Poleg ovrednotenja uspešnosti posameznih metod v nalogi predstavimo tudi razmišljanje o tem, zakaj je zaznavanje čustvene naravnanosti v besedilih veliko težji problem kot zaznavanje teme, ki jo je možno razbrati že iz posameznih ključnih besed. Ker smo metode preizkusili na različnih predstavitvah podatkov, bomo uspešnost ovrednotili tudi s statistično primerjavo le-teh med seboj.

Poglavje 2

Problemska domena in podatki

Za potrebe naše raziskave smo se odločili, da bomo v podrobnogled vzeli komentarje, ki so razvrščeni pod določenim člankom, objavljenim na znanem slovenskem spletnem portalu RTV-SLO[1]. Izbrali smo si članek z naslovom "FDV: Magistrsko delo premierke Bratušek ni plagiat", ki je bil objavljen 1. julija 2013[2]. Objavljeni članek je bil obsežno komentiran in smo zanj predvidevali, da se bodo v komentarjih pod člankom kresala različna mnenja. Po bližnji seznanitvi z vsebino članka in komentarjev smo ugotovili, da je to primerno gradivo za preučevanje naših pristopov, saj je bilo komentarjev precej več kot pri ostalih člankih, ki smo jih zasledili, ti pa so v dovolj veliki meri zastopali oba razreda, ki ju bomo podrobneje opisali kasneje. V nadaljevanju sledi obdelava komentarjev do te mere, da smo na njih lahko izvajali različne matematične operacije.

[1] `http://www.rtvslo.si/`

[2] `http://www.rtvslo.si/slovenija/fdv-magistrsko-delo-premierke-bratusek-ni-plagiat/312209`

2.1 Zajem podatkov

Besedilne podatke smo pred obdelavo s tehnikami strojnega učenja morali primerno predstaviti in jih zapisati v obliki, ki je primerna za izbrane tehnike. Za strojno učenje smo v diplomski nalogi izbrali metode, ki uporabljajo atributne zapise vhodnih podatkov. Naše podatke tako predstavlja matrika, ki jo sestavljajo vrstice (primeri) in stolpci (atributi) in ki za dani atribut in primer v matriki vsebujejo določeno numerično vrednost. Vsaka vrstica v učnih podatkih vsebuje tudi razred primera.

Članek, ki smo ga izbrali za analizo, je bilo potrebno najprej prebrati, da smo se seznanili s temo, na katero se bodo komentarji nanašali. Nato smo vsak komentar, prikazan pod člankom, ročno po lastni presoji razvrstili kot pozitiven ("poz" oz. 1) oz. negativen ("neg" oz. 0) glede na to, ali avtor komentarja izraža strinjanje oz. nestrinjanje z napisanim v članku. Če se osredotočimo na tri glavne načine za razvrščanje primerov glede na mnenje, ki jih v svojem prispevku opisujeta Kim in Hovy [14] - besedna raven, povedna raven in dokumentna raven - lahko rečemo, da naša raziskava bazira na dokumentni ravni, vendar pa primeri v naših podatkih včasih predstavljajo le posamezne besede, en stavek ali pa odstavek.

V naši problemski domeni torej primere izvorno predstavljajo različno dolga besedila, ki smo jih ročno uvrstili v dva razreda. V tem prvem koraku predobdelave podatkov je nastala tekstovna datoteka v spodnji obliki:

```
Razred Besedilo
poz A je sedaj g. Tanko zadovoljen, ali ga še kaj muči glede...
poz :D
neg ..hahahhaha,..hahahhaa,...fdv....hahah..
poz upam, da se bo zdej nehalo s temi preverbami...
poz Pričakujem cel kup komentator o tem, da je FDV pod polit...
neg in potem svizec zavije čokolado, sloni letijo in obstaja...
```

```
neg Vrana vrani ne izkljuje oči ... Sramota za rdečo fakulteto.
```

Sledila je obdelava komentarjev. Iz besedil smo odstranili vse znake in ločila, da smo na koncu dobili poljubno dolge nize črk. Za odstranitev vseh znakov razen črk smo se odločili, ker želimo čustveno naravnanost avtorjev v besedilih odkriti le na podlagi besed, ki bi kazale na določen sentiment in bi se lahko pokazale pri tvorjenju n-terk. Po poglobitvi v različno literaturo, ki se nanaša na temo našega problema, smo se odločili, da k rešitvi pristopimo na malo drugačen, bolj tehnični način, z manj ozira na semantiko in slovnične zahteve slovenskega jezika. Značilke so v našem primeru n-terke zaporednih črk. Ker optimalne dolžine zaporednih črk atributov nismo poznali, smo vse napovedne modele preizkusili za $n = 2 \ldots 8$ znakov. Za ta razpon števila znakov smo se odločili na podlagi tega, da je povprečna dolžina vseh besed v izbranih slovenskih leposlovnih besedilih $4,5$ črk, povprečna dolžina različnih besed v istih slovenskih leposlovnih besedilih pa 8 črk (Vodopivec [15]). Trojice in dvojice črk so bile v analizo dodane, da bi pokazali razliko med informativnostjo samih atributov, torej kako dolžina niza črk vpliva na samo značilnost nekega atributa za dani razred oz. kako dolžina niza črk pripomore k boljšemu učenju metode na učnih podatkih.

Glede na zgoraj zapisano, je vrednost atributov za dani komentar enaka število ponovitev dotične n-terke črk v danem primeru. Torej za vsak komentar štejemo, kolikokrat se katera izmed n-terk v nizu črk ponovi, to pa predstavlja eno vrstico v naših podatkih. Ker so komentarji različno dolgi, je bilo potrebno vse vrstice normalizirati. Na koncu torej vrednosti atributov predstavljajo deleže zastopanosti teh atributov v primeru, oziroma atributi predstavljajo relativno frekvenco dane n-terke v komentarju.

2.2 Predstavitev podatkov

Sedaj nam je torej znana struktura podatkov, nad katerimi bomo izvedli strojno učenje. Ker pa nas bo v nalogi zanimalo predvsem, kaj nam ti podatki sploh povedo oziroma česa se iz njih lahko naučimo, je prav, da predstavimo nekaj ključnih dejstev, na podlagi katerih bomo lažje potegnili sklepne ugotovitve.

Članek, ki smo ga vzeli pod drobnogled, ima 540 komentarjev, kar pomeni, da ima naša podatkovna matrika 540 vrstic. Vsak komentar je en primer oz. vrstica.

Komentarji oz. nizi črk so bili različno dolgi. Najkrajšega predstavlja le ena črka, najdaljšega pa kar 2426 črk. Povprečna dolžina enega niza črk znaša 195 znakov in predstavlja mejo med 34% komentarjev, ki so daljši od povprečne dolžine, in 66% komentarji, ki so od povprečne dolžine krajši.

Vsakega od komentarjev smo ročno razvrstili v od enega od razredov ("poz"in "neg"), ti pa so v celotnem naboru podatkov zastopani v razmerju poz:neg = 4:6; 40% komentarjev je bilo torej spoznanih za pozitivne.

Glede na to, da smo nabore atributov določili za sedem naborov n-terk ($n = 2 \ldots 8$), smo zato zgradili sedem različnih podatkovih matrik. Nekatere njihove statistične lastnosti predstavimo v tabeli 2.1.

Iz tabele je razvidno, da gre za redko porazdeljene matrike podatkov. To je razvidno predvsem pri matrikah, kjer atribute predstavljajo terke z vsaj 4 črkami. Daljša kot je n-terka, manjša bo verjetnost, da bo specifična kombinacija črk zastopane tudi v besedilu komentarja.

2.3 Podporni podatki

Da bi pokazali in podprli trditev, da se tehnike strojnega učenja, ki smo jih izbrali za naš problem, za razvrščanje besedilnih podatkov sicer zelo dobro obnesejo, vendar pa zaradi nekaterih dejavnikov niso pokazale dobrih rezul-

Tabela 2.1: Zastopanost atributov v matrikah podatkov glede na različne dolžine terk

	n=2	n=3	n=4	n=5	n=6	n=7	n=8
Št. vseh značilk (tiste, ki so prisotne v vsaj 1 primeru)	627	5892	24606	50432	68412	78753	84790
Št. vseh značilk, ki so prisotne v vsaj 2 primerih	571	4478	13406	17122	15184	12723	10918
Št. vseh značilk, ki so prisotne v vsaj 6 primerih	487	2717	4423	2573	1351	797	507
Št. vseh značilk, ki so prisotne v vsaj 11 primerih	440	1891	1980	841	393	211	117

tatov, smo ustvarili podobno problemsko domeno, kjer smo se s tehničnega vidika želeli kar najbolj približati dejanskim podatkom.

Tokrat smo s portala RTV-SLO vzeli 400 člankov in jih razvrstili glede na žanre (teme). V novi problemski domeni so torej naši primeri namesto komentarjev, ki se nanašajo na neko dotično tematiko, članki, ki se navezujejo na določeno temo. Samo število primerov je tu sicer nekoliko manjše od tistega pri komentarjih, vendar bomo v kasnejših poglavjih pokazali, da je bilo za ta eksperiment zajetih dovolj podatkov, da smo lahko dokazali naše domneve.

Da bi zajeli podoben aspekt pripisovanja primerov določenim razredom, smo se tudi tu odločili, da zajamemo članke iz dveh različnih tem. Razreda "poz"oz. "neg"tukaj zamenjata razreda "šport"in "novice". Razred primera smo določili skladno z zavihkom spletne strani, pod katerim so bili članki razvrščeni na spletnem portalu (šport, novice). Na tem mestu lahko omenimo že prvo bistveno razliko, ki je na prvi pogled med tema dvema domenama morda ne bi opazili. Gre namreč za to, da smo komentarje, kot je opisano v prejšnjem podpoglavju, razvrstili glede na lastno subjektivno oceno, ki je bila zasnova na podlagi ene osebe. Pri razvrstitvi člankov v različna žanra pa smo se izognili le enemu samemu mnenju, saj nam ni bilo potrebno oceniti, v kateri žaner nek članek spada (za to so poskrbeli že pisci besedil, ki so svoje članke razvrstili v primeren zavihek na strani).

Ko so bili primeri dodeljeni različnima razredoma, je sledila enaka obdelava besedila kot pri razvrščanju komentarjev. Tudi tokrat so bili seveda nizi črk različno dolgi, vendar v primerjavi s komentarji precej daljši. Najkrajše besedilo predstavlja 294 črk, najdaljšega pa kar 2426 črk. V povprečju so imeli članki 2162 črk, kar je približno 11-krat več kot pri komentarjih. Povprečje tu predstavlja mejo med 40% komentarjev, ki so daljši od povprečne dolžine, in 60% komentarji, ki so od povprečne dolžine krajši.

Ker smo želeli podatke čimbolj približali tistim, ki smo jih pridobili na

komentarjih, smo se omejili na prvih 195 črk vsakega članka. Tako smo prišli do povprečne dolžine črkovnega zaporedja, ki je 195 črk, kar je ravno povprečna dolžina niza črk pri komentarjih.

Razmerje med izbranima razredoma je v tem primeru enakomerno porazdeljeno, za vsak žanr smo namreč opredelili 50% od vseh primerov. V tem pogledu se tudi to razmerje nekoliko razlikuje od tistega pri komentarjih, vendar je odstopanje majhno.

Seveda smo tudi v tem primeru atribute določili kot n-terke v že znanem razponu števila črk n, lastnosti dobljenih podatkovnih matrik pa za lažjo primerjavo za tako dobljene podatke predstavljamo v tabeli 2.2. Tudi iz te tabele lahko povzamemo, da gre za podobne podatke kot pri analizi komentarjev.

Tabela 2.2: Zastopanost atributov v matrikah podatkov glede na različne dolžine terk pri člankih

	n=2	n=3	n=4	n=5	n=6	n=7	n=8
Št. vseh značilk (tiste, ki so prisotne v vsaj 1 primeru)	644	5937	22845	41764	53353	59919	63964
Št. vseh značilk, ki so prisotne v vsaj 2 primerih	594	4374	10896	11056	8412	6228	4642
Št. vseh značilk, ki so prisotne v vsaj 6 primerih	495	2413	2819	1393	719	414	263
Št. vseh značilk, ki so prisotne v vsaj 11 primerih	431	1599	1130	423	215	122	77

Poglavje 3

Metode

V tem poglavju se bomo osredotočili na predstavitev uporabljenih tehnik strojnega učenja in pristopov k ocenjevanju njihove napovedne točnosti. Opisali bomo, kako metode delujejo in zakaj so prav te pomembne pri iskanju odgovorov na vprašanja, ki se pojavljajo v zvezi s to tematiko.

3.1 Napovedni modeli

S tehnikami strojnega učenja lahko iz učnih podatkov gradimo klasifikacijske napovedne modele, ki na podlagi atributnega opisa testnega primera tega razvrstijo v enega od ciljnih razredov. V naši problemski domeni je bil učni problem dvorazredni, problem pa klasifikacija v razred 0 oz. 1 (”poz”in ”neg”oziroma ”novice”in ”šport”). V nalogi smo preizkusili štiri dobro poznane in uveljavljene metode, dodatno pa skušali napovedno točnost izboljšati s tehniko ansambla klasifikatorjev.

3.1.1 Logistična regresija

Logistična regresija se uporablja za napovedovanje izida kategorično odvisne spremenljivke (razreda) na osnovi ene ali več neodvisnih spremenljivk

(atributov). Verjetnosti, ki jih dobimo z uporabo logistične funkcije, opisujejo možne izide glede na dano kombinacijo atributov. Logistična regresija se lahko nanaša na problem, v katerem je odvisna spremenljivka binarna – to pomeni, da imamo dva možna razreda – ali pa imamo na voljo več razredov, ki jih lahko pripišemo dani kombinaciji značilk. V našem primeru uporabljamo binarni razred, saj vsakega od primerov lahko klasificiramo kot pozitivnega oz. negativnega (pri člankih ali se nanša na šport ali na novice), torej razred zavzema natančno dve vrednosti.

Logistično regresijo [4] smo implementirali sami na podlagi predavanj Andrewa Nga[1] in si pri tem pomagali s knjižnico scipy[2], iz katere je bil za optimizacijo vzet algoritem L-BFGS.

Za grajanje modela logistčne regresije moramo najprej opredeliti funkcijo hipoteze, ki vrne vrednosti med 0 in 1, saj napovedujemo dva možna razreda. Ta je predstavljena s formulo

$$h_\theta(x) = g(\theta^T x) = P(y = 1|x; \theta) \tag{3.1}$$

in vrne vrjetnost za $y = 1$ oz, da primeru x pripada razred y.

Funkcijo hipoteze izračunamo z logistično funkcijo g in tako zagotovimo, da bodo napovedne verjetnosti zavzemale vrednosti med 0 in 1. Sigmoidna funcija je predstavljena s formulo

$$g(z) = \frac{1}{1 + e^{-z}}. \tag{3.2}$$

Torej cilj modeliranja je iskanje takega parametra θ, da bomo prišli do čimbolj natančne vrednosti funkcije hipoteze oziroma do čimbolj točne napovedi.

[1] `http://openclassroom.stanford.edu/MainFolder/VideoPage.php?course=MachineLearning&video=04.1-LogisticRegression-Classification&speed=100`

[2] `http://www.scipy.org/`

Podobno kot pri linearni regresiji, tudi tukaj opredelimo cenovno funkcijo, ki bo podala oceno napake funkcije hipoteze, ki se pri dani vrednosti θ prilega našim podatkom. Cenovna funkcija je prilagojena za regularizirano (λ večja od 0) in neregularizirano ($\lambda = 0$) logistično regresijo. Regularizacijski parameter bo na koncu vodil k manjšim vrednostim theta, s čimer se izognemo prevelikemu prileganju tesnim podatkov za napovedovanje razreda. Cenovna funkcija je predstavljena s formulo:

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m}\left[y^{(i)}\log(h_\theta(x^{(i)}) + (1-y^{(i)})\log(1-h_\theta(x^{(i)})\right] + \frac{\lambda}{2}\sum_{j=1}^{n}\theta_j^2. \tag{3.3}$$

Če želimo ugotoviti najprimernejšo funkcijo hipoteze, moramo najti tako vrednost θ, ki zmanjšuje vrednost $J(\theta)$. To je mogoče doseči z iskanjem gradienta cenovne funkcije. Parameter θ ponovno izračunamo na spodnji način

$$\theta^{(t+1)} = \theta^{(t)} - H^{-1}\nabla_\theta J, \tag{3.4}$$

kjer je gradientna funkcija predstavljena s formulo:

$$\nabla_\theta J = \frac{1}{m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})x^{(i)} + \frac{\lambda}{m}\theta. \tag{3.5}$$

3.1.2 Metoda podpornih vektorjev

Metoda podpornih vektorjev zagotavlja uveljavljen in učinkovit način razvrščanja za analizo podatkov in iskanje najmanj tvegane ločitve med različnimi razredi. Iskanje meje med razredoma je močno odvisna od razpoložljivega nabora podatkov in pa optimizacijskih parametrov. Tehnike za izbor najboljših atributov in SVM optimizacija parametrov sta v kombinaciji znana po tem, da izboljšata natančnost klasifikacije.

Za učinkovito razvrstitev podatkov mora SVM najprej poiskati maksimalno mejo, ki loči dva razreda, nato pa postaviti separator s hiperravnino, ki bo ločila primere, ki se klasificirajo v en ali drug razred. Novi podatki so razvrščeni po odoločitvi, na katero stran hiperravnine spadajo, s tem pa je odločeno, kateremu razredu so bili dodeljeni. Vendar pa nekateri vhodni prostori niso dovolj dobro ločljivi v linearni ravnini, zato se pogosto uporabljajo preslikave vhodnega prostora v višje dimenzionalni prostor, kjer primere lahko lažje ločimo. Razdaljo vektorjev, ki ležijo najbližje hiperravnini, pri tem maksimiramo, saj želimo ustvariti čimbolj eksplicitne odločitve tudi za primere, ki niso čisto enaki tistim, na katerih smo se učili. Za implementacijo SVM modela smo se poslužili knjižnice sklearn[3].

SVM opišemo z množico primerov:

$$\{(x_1, y_1), ..., (x_m, y_m)\}, x_i \in X, y_i \in \{-1, 1\} \tag{3.6}$$

y_i tako predstavlja razred pripadajočemu primeru x_i.

Klasifikator nato izračuna hiperravnino, ki množice primerov obeh razredov loči karseda najbolje. Ta ravnina je podana z normalnim vektorjem w in pragom b.

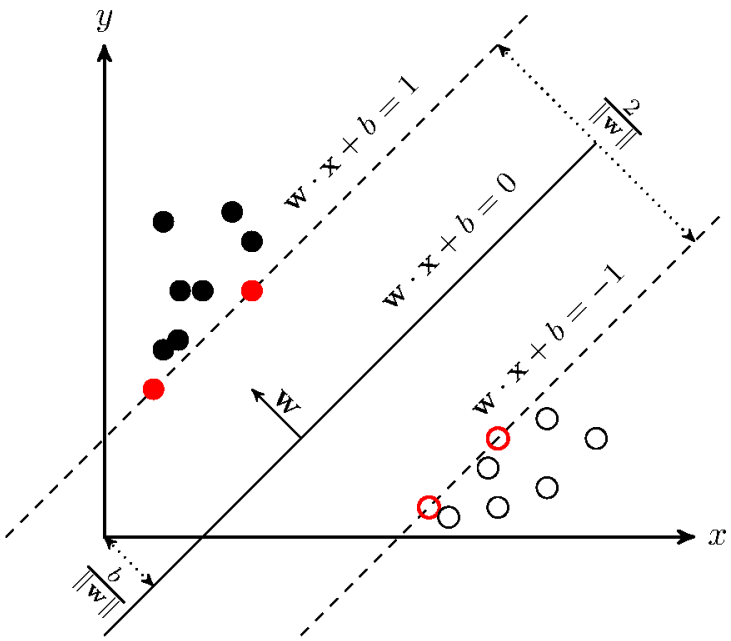
Za primer x_i iz učne množice se pri tem priredi predznak odločitvene funkcije:

$$y_i = sgn(\langle w, x_i \rangle + b) \tag{3.7}$$

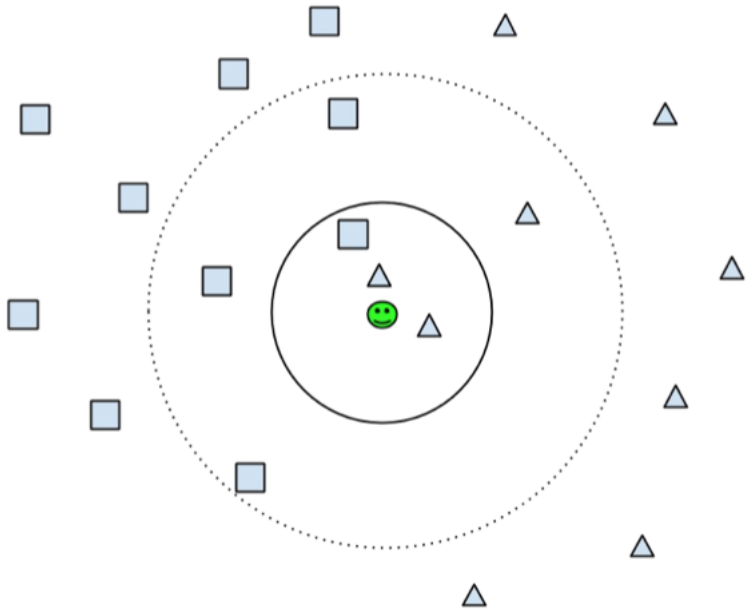
Rezultat je lahko pozitiven ali negativen in je odvisen od tega, ali se določen primer nahaja na eni ali drugi strani hiperravnine. Za boljšo predstavo si to poglejmo na sliki 3.1[4].

[3] `http://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html`

[4] `http://www.pengyifan.com/blog/wp-content/uploads/2013/09/svm.png`



Slika 3.1: Prikaz ločitve primerov v dva razreda s hiperravnino.



Slika 3.2: Prikaz ugotavljanja najbljižjih sosedov.

3.1.3 Metoda k najbljižjih sosedov

Metoda k najbližjih sosedov za osnovo vzame kar učne primere same. Ko mora za novi primer določiti, v kateri razred ga bo potrebno klasificirati, ta klasifikacijska tehnika poišče v učni množici k takih primerov, ki so novemu primeru najbolj podobni. Rezultat napovedi je verjetnostna porazdelitev števila primerov, ki pripadajo posameznim razredom v množici k najbolj podobnih primerov. Za boljše razumevanje si poglejmo predstavitev algoritma na sliki 3.2[5]

V našem primeru smo uporabili metodo KNN iz knjižnice sklearn[6]. Podobnost med primeri smo ocenjevali z Evklidsko razdaljo.

[5] `https://jeena.net/images/2013/catdog/k-nearest-neighbours.png`

[6] `http://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html`

Slaba lastnost večine glasov, ki klasifikacirajo nek primer v enega izmed razredov, se pojavi, ko je porazdelitev razreda popačena. To pomeni, da pogostejši razred prevladuje pri napovedi novega primera zato, ker so po navadi pogosti med k najbližjimi sosedi zaradi njihovega velikega števila. To lahko rešimo tako, da je vsak izmed k najbližjih sosedov nekega primera utežen z nekim številom točk, upoštevajoč razdaljo med preskusno točko za vsako od svojih k bližnjih sosedov. Razred vsake izmed k najbližjih točk se pomnoži s težo obratnosorazmerne oddaljenosti od te točke do preskusnih točk. V našem primeru smo to upoštevali s parametrom weights='distance'.

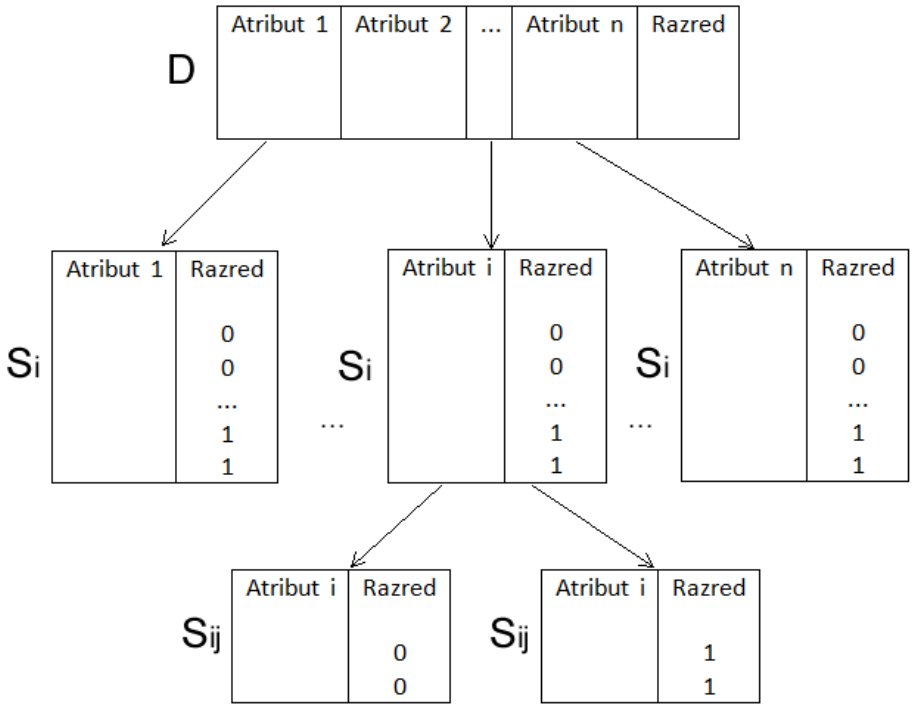
3.1.4 Naključni gozdovi

Za razlago metode naključnih gozdov moramo najprej spoznati strukturo in zgradbo enega klasifikacijskega drevesa. Le-to ima hierarhično obliko, ki se uporablja za razvrščanje razredov glede na vrsto vprašanj ali pravil, ki se nanašajo na atribute določenega razreda. Atributi razredov so lahko vse spremenljivke z binarno, nominalno, ordinalno in kvantitativno vrednostjo.

Prvi korak gradnje drevesa je izračun verjetnosti za pojavitev nekega razreda p_j. Upoštevati moramo, da se, ko računamo verjetnosti, osredotočimo le na pojavnost razredov in ne na njihove atribute. Ko poznamo verjetnosti pojavitve posameznih razredov, lahko izračunamo stopnjo čistosti posameznih tabel z eno od treh meram, ki nam bodo pomagale pri gradnji klasifikacijskega drevesa. To so entropija, Gini indeks in klasifikacijska napaka. V našem primeru smo uporabili Gini indeks po spodnji formuli:

$$Gini = 1 - \sum_j p_j^2 \tag{3.8}$$

Če v podatkih obstaja le en razred, Gini indeks zasede vrednost 0, saj je verjetnost pojavitve razreda enaka 1. Gini indeks prav tako doseže svoj maksimum, ko imajo vsi razredi v podatkih enake verjetnosti $p = 1/n$, ve-



Slika 3.3: Potek razvejevanja in računanja stopnje čistosti.

dno pa zasede vrednost med 0 in 1 ne glede na število različnih razredov v podatkih.

Naši podatki so predstavljeni v tabeli D z atrubuti in pripadajočimi razredi. Iz tabele D vzamemo vsak atribut posebej z njegovimi pripadajočimi razredi in tako ustvarimo podtabele S_i. Kolikor je različnih atributov, toliko je tudi novih podtabel S_i. Za vse elemente v strukturi nato izračunamo vrednosti entropije, Gini indeksa in klasifikacijske napake. Za boljšo predstavo si oglejmo potek na sliki 3.3.

Različne načine računanja čistosti tabele D in podtabel S_i uporabimo zato, da primerjamo razlike v stopnji čistosti med njimi preden jih razdelimo na več delov. Za mero, s katero primerjamo razlike v čistosti tabel, uporabimo informacijo I. Zanima nas, kakšno informacijo dobimo, če tabelo podatkov

razdelimo glede na vrednosti atributov. To izračunamo po spodnji formuli:

$$I_i = Gini_D - \sum_j \frac{k}{n} \cdot Gini_{S_{ij}} \tag{3.9}$$

Spremenljivka k predstavlja število primerov v podtabeli S_{ij}, n pa število vseh primerov v tabeli D. Za vsak atribut v tabeli D tako izračunamo informacijo in nato izberemo atribut, pri katerem je bila ta največja:

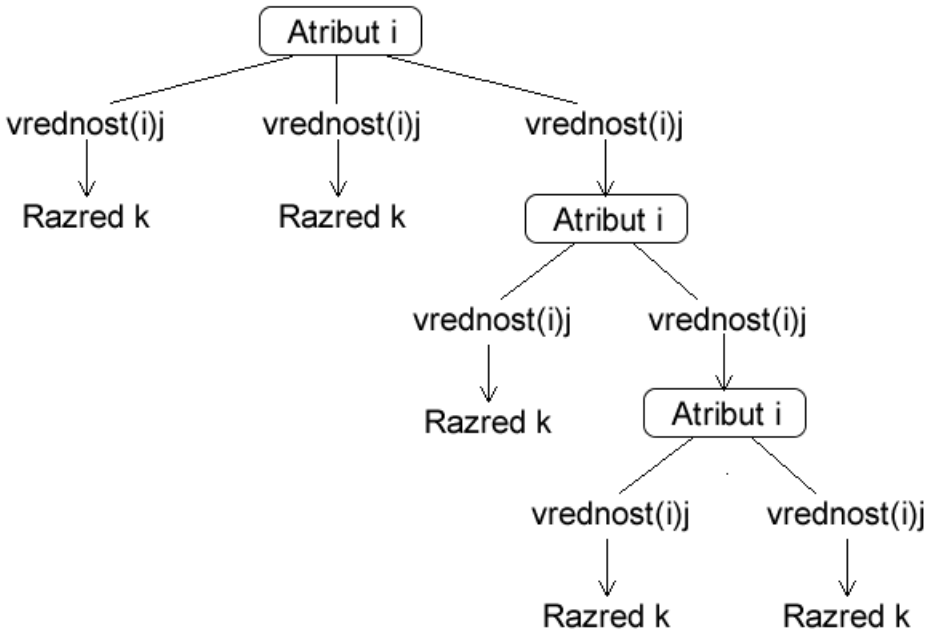
$$i = argmax\left\{I_i\right\}. \tag{3.10}$$

Izbrani atribut i nato postane vozlišče (v prvi iteraciji koren) v odločitvenem drevesu, tabelo D pa razdelimo v podtabele glede na vrednosti atributa i. Nato postopek ponavljamo, dokler ne pridemo do listov odločitvenega modela, ki je prikazan na sliki 3.4.

Ko je odločitveno drevo oblikovano, lahko vsakemu naslednjemu primeru napovemo razred, tako da glede na pravila v drevesu in vrednosti atributov pridemo do lista, ki predstavlja razred.

Sedaj, ko poznamo postopek gradnje enega klasifikacijskega drevesa, pa razložimo še metodo naključnih gozdov [9]. Ta namesto enega klasifikacijskega drevesa upošteva kar množico oziroma l takih dreves. Za razvrščanje novega primera, je vhodni podatek za vseh l dreves prav vhodni vektor. Vsako drevo iz gozda nato poda svojo napoved - oceno o tem, v kateri razred primer spada. Naključni gozd primer razvrsti v razred, ki ga je napovedala večina klasifikacijskih dreves v gozdu. Za potrebe ocene verjetnosti razredov pa so te izračunane iz števila dreves, ki glasujejo za posamezen razred.

Kakovost naključnih gozdov temelji na raznolikosti dreves. Da dosežemo to raznolikost, učne primere za posamezno drevo vzorčimo (s ponovitvami) iz učne množice tako, da je vzorec enako velik kot učna množica. Če je v podatkih M vhodnih atributov, število $m << M$ določimo naključno iz M tako, da ta kar najbolje razdeli množico M. Vrednost m predstavlja število



Slika 3.4: Gradnja klasifikacijskega drevesa na podlagi računanja informacije in deljenja tabel.

atributov, ki jih upoštevamo pri gradnji klasifikacijskih dreves in je v našem primeru $\sqrt{M}$.

Za implementacijo metode naključnih gozdov smo se prav tako poslužili knjižnice sklearn[7].

3.1.5 Skladanje

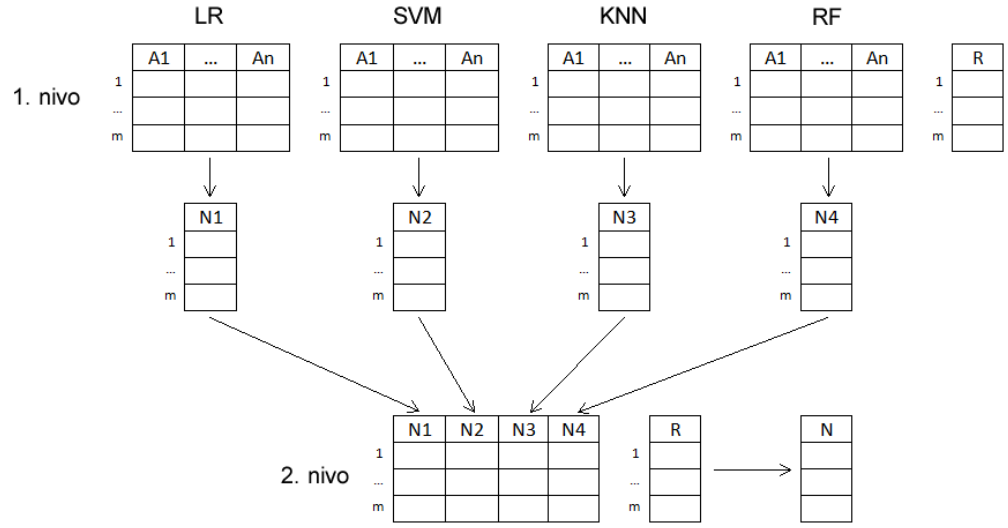
Za posamezne metode smo kaj kmalu ugotovili, kakšne ocene lahko dosežejo, dodatno pa nas je zanimalo, ali lahko napovedi posameznih razredov primerom še izboljšamo z združevanjem različnih pristopov. Za tehniko združevanja smo izbrali in implementirali metodo skladanja tako, ko jo je predlagal Wolpert [6].

V namene združevanja verjetnosti ocen različnih klasifikatorjev smo s prečnim preverjanjem na učni množici za vsak primer izračunali verjetnosti razredov z uporabo vseh štirih klasifikatorjev, torej z uporabo logistične regresije, metodo podpornih vektorjev, k najbližjih sosedov in naključnega gozda dreves.

Metoda skladanja deluje v dveh korakih. Na prvem nivoji metode učenja podajo svoje napovedi, na drugem nivoju pa te napovedi združimo v novo matriko podatkov in na njej poženemo metodo učenja, ki bo podala končne napovedi.

Vsak od prej omenjenih štirih klasifikatorjev torej služi kot učenec na prvem nivoju in pri prečnem preverjanju vrne vektor napovedi. Tako smo za vsako predstavitev podatkov dobili štiri vektorje, ti pa na drugem nivoju predstavljajo stolpce v novi matriki učenja. Resnične vrednosti razredov za vsak primer iz podatkov ostajajo nespremenjene, paziti pa je potrebno tudi na to, da pri prečnem preverjanju z vsemi klasifikatorji podamo res pravo napoved za dotični primer, torej da se originalna vrednost razreda ne izgubi

[7] `http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html`



Slika 3.5: Gradnja nove tabele za učenje na drugem nivoju pri metodi skladanja.

ali pomeša. Za boljšo predstavo poglejmo še potek skladanja na sliki 3.5.

Ker so novi podatki med seboj neodvisni, smo za učenca na drugem nivoju uporabili logistično regresijo, ki bo vrnila končne napovedi. Upamo, da bodo nove napovedi morebiti boljše od tistih, ki so jih podale metode na prvem nivoju. Ta postopek smo izvedli za vsak set podatkov posebej, torej sedemkrat. Zaradi velike časovne zahtevnosti celotnega postopka in priprave nove tabele smo na prvem nivoju za vsak klasifikator izvedli 10-kratno prečno preverjanje, da smo pridobili stolpce za novo matriko, na drugem nivoju pa 5-kratno prečno preverjanje, saj smo se želeli izogniti naključnemu rezultatu te metode.

3.2 Ocenjevanje kakovosti napovednih modelov

Točnost napovednih modelov smo ocenili z tehniko prečnega preverjanja. Učno množico smo razdelili na 10 približno enakih množic, in potem v vsaki od deset iteracij eno od teh izbrali za testiranje klasifikatorjev ki so bili zgrajeni na primerih iz devetih preostalih množic. Pri tem smo uporabili dve metrike ocenjevanja točnosti, ki jih opišemo spodaj. V rezultatih podajamo njihove povprečne vrednosti preko desetih iteracij učenja in testiranja.

3.2.1 Ocena LogLoss

Klasifikacijski modeli kot rezultat svojih napovedi vrnejo verjetnost pripadnosti razredom. Navadno to pomeni, da nobena napovedna metoda ni 100-odstotno prepričana v svoje napovedi (razen, če se v učni množici niso ponovili isti primeri kot v testni), vedno obstaja delež, ki dopušča možnost napake. Logaritem funkcije verjetja za Bernoullijevo naključno porazdelitev se uporablja za oceno napake, ki jo napravimo pri napovedovanju tega, s kolikšno verjetnostjo nekaj drži ali ne, kjer 1 pomeni popolni zadetek, 0 pa zgrešene napovedi. Ocena LogLoss torej pove, kako odločen je bil napovedni model pri svojih napovedih, to pa naredi tako, da najbolj kaznuje tiste napovedi, pri katerih smo najbolj zgrešili. Izračunamo jo po naslednji enačbi:

$$\text{LogLoss} = -\frac{1}{N}\sum_{i=1}^{N} y_i \log\left(\hat{y}_i\right) + \left(1 - y_i\right) \log\left(1 - \hat{y}_i\right), \tag{3.11}$$

kjer je N število primerov, log naravni algoritem, $\hat{y}_i$ verjetnost napovedi čustvene naravananosti avtorja besedila na možne n-terke za i-ti primer in y_i prava vrednost razreda pri i-tem primeru ($y_i = 1$, $y_i = 0$). Manjša kot je ocena logloss, boljši je model, s katerim podamo svoje napovedi.

3.2.2 Delež pravilno razvrščenih primerov

Druga ocena, ki smo jo uporabili pri meritvi uspešnosti napovednih modelov, je v literaturi veliko bolj pogosta. Prisotna je v večini primerih, kjer govorimo o metodah, ki na podlagi statističnih principov napovedujejo prisotnosti razredov v primerih. Ocenjevanje uspešnosti modela smo izvedli z izračunom točnosti. Za vsak primer iz množice smo primerjali napovedano vrednost z dejansko. Ocena nam poroča o deležu primerov, za katerih so bile napovedi klasifikacijskega modela pravilne in je predstavljena s spodnjo formulo, kjer SPN pomeni število pravilno napovedanih, SVP pa število vseh primerov.

$$\mathrm{t}ocnost = \frac{SPN}{SVP} \tag{3.12}$$

Poglavje 4

Izbira parametrov učenja

Uporabljene metode učenja, ki smo jih v nalogi uporabili, so odvisne od za metodo specifičnih parametrov. Od vrednosti teh parametrov je lahko odvisna točnost napovedi modelov, ki jo dobimo z tehniko strojnega učenja. Ustrezne vrednosti parametrov ocenimo tehniko internega prečnega preverjanja na samo učnih primerih. Za dani nabor možnih vrednosti parametrov to storimo s 5-kratnem prečnim preverjanjem na učni množici, ter potem za gradnjo klasifikatorja na celotni učni množici uporabimo vrednost parametra, pri katerem smo dosegli najvišjo povprečno točnost. Spodaj predstavimo še razpon parametrov za vsako od uporabljenih metod.

4.1 Logistična regresija

Logistična regresija pri svojem računanju potrebuje regularizacijski parameter λ, zato želimo vedeti, kako izbrati najboljšo vrednost le-tega. Za izbiro optimalnega parametra s pomočjo prečnega preverjanja smo preizkusili več različnih vrednosti λ:

$$\lambda = \left\{ \begin{array}{c} 1e-01, 1e-02, 1e-03, 1e-04, 1e-05, \\ 1e-06, 1e-07, 1e-08, 1e-09, 1e-10 \end{array} \right\} \tag{4.1}$$

Optimalen parameter λ vodi do manjše vrednosti θ, to pa preprečuje, da bi se metoda preveč prilagodila učnim podatkom.

4.2 Metoda podpornih vektorjev

Glede na prisotnost parametrov pri SVM, ki lahko vplivajo na izid učenja in klasifikacije, je povsem logično, da to lahko izboljša učinkovitost algoritma. Najbolj osnovni pristop pri SVM za izboljšanje razvrščanja je kontrola uteži, ki jo kaznujemo s parametrom c (s parametrom c množimo uteži) in iskanje najboljšega kompromisa med nezaznanimi napakami in posplošitvijo modela. Pri naši problemski domeni smo ugotovili, da je smiselno preveriti naslednji nabor parametrov:

$$c = \{200, 300, 400, 500, 600, 700\} \tag{4.2}$$

Visoke vrednosti parametra c, bodo v veliki meri kaznovale napačno obravnave primere, zato bo posledično hiperravnina tista, kjer se bomo izognili napakam razvrščanja. Če pa parameter c zaseda nizke vrednosti in tako le rahlo kaznuje napačne klasifikacije, je rezultat lahko napačna ločitev primerov v en in drugi razred (Gaspar in drugi [11]).

4.3 Metoda k najbljižjih sosedov

Najboljša izbira parametra k je odvisna od podatkov samih. Na splošno večje vrednosti k vplivajo na zmanjšanje šuma v klasifikaciji, vendar meje med razredi niso več tako zelo jasne. Natančnost KNN algoritma se lahko močno slabša zaradi prisotnosti nepomembnih značilnosti ali če lestvice značilnosti niso v skladu z njihovo pomembnostjo (Han [16]). Pri tem algoritmu pričakujemo, da se bodo bolje odrezale nizke k vrednosti, saj večanje le-the naredi

ves sistem bolj kompleksen. Mi smo optimalno vrednost k iskali v naslednjem naboru vrednosti:

$$k = \{4, 6, 8, 10, 12, 13, 20, 30, 40, 50, 100\} \tag{4.3}$$

V binarnih klasifikacijski problem, kot je naš, je koristno, da za parameter k izberemo liho število, saj se s tem izognemo izenačenim primerom, vendar pa to ni nujno.

4.4 Naključni gozdovi

Glavno načelo metode naključnih gozdov je, da lahko skupina klasifikacijskih dreves skupaj tvori močen model za učenje. Vsak klasifikator je posamezno veliko slabši učenec od skupine večih klasifikatorjev, ki delujejo znotraj gozda vsak posamezno, navzven pa homogeno. Za naš nabor podatkov smo preizkusili naslednji razpon velikosti gozda:

$$n = \{100, 150, 200, 250, 300, 350\} \tag{4.4}$$

Pri izbiri optimalnega števila dreves n v gozdu moramo upoštevati predvsem to, da večji kot je gozd, veča se do neke mere uspešnost metode, vendar pa se tudi kompleksnost in časovna zahtevnost algoritma hitro povečujeta.

Poglavje 5

Rezultati in vrednotenje

V tem poglavju bomo predstavili, kakšne rezultate so dale metode na različnih naborih podatkov, kateri podatki so za naš problem najbolj primerni, kateri parametri so privedli metode do takšnih vrednosti ocen točnosti in logloss in kaj je botrovalo k takšnim končnim rezultatom. V nadaljnih točkah tega poglavja bomo predstavili strnjene ugotovite, podrobnejši rezultati pa se nahajajo v prilogah A, B, C, D in E.

5.1 Napovedna točnost

Glede na oceno logloss pri preučevanju komentarjev smo prišli do rezultatov, ki so prikazani v tabeli 5.1. Najboljšo točnost je dosegla metoda naključnih gozdov na podatkih, kjer so atribute predstavljale trojke črk, in sicer vrednost 0.623. Temu rezultatu sledita še logistična regresija z oceno 0.624 in metoda podpornih vektorjev z oceno 0.627.

Zanimivo je, da smo najslabši rezultat dobili prav tako pri metodi naključnih gozdov na podatkih, kjer so značilke predstavljale sedmerke črk, in sicer 0.744. Iz tega lahko razberemo, da je metoda naključnih gozdov najbolj občutljiva na to, kako so predstavljeni podatki. Za drugo najslabšo metodo pa se je izkazala metoda k najbljižjih sosedov z oceno 0.697, ki je bila iz-

merjena na osmerkah črk. Prav vse metode so bile najbolje ocenjene na podatkih, predstavljenih s trojkami črk, najslabše pa na podatkih iz osmerk črk, z izjemo metode naključnih gozdov, ki je najslabšo oceno dosegla pri podatkih s sedmerkami črk.

Tabela 5.1: Rezultati metod glede na oceno logloss pri razvrščanju komentarjev.

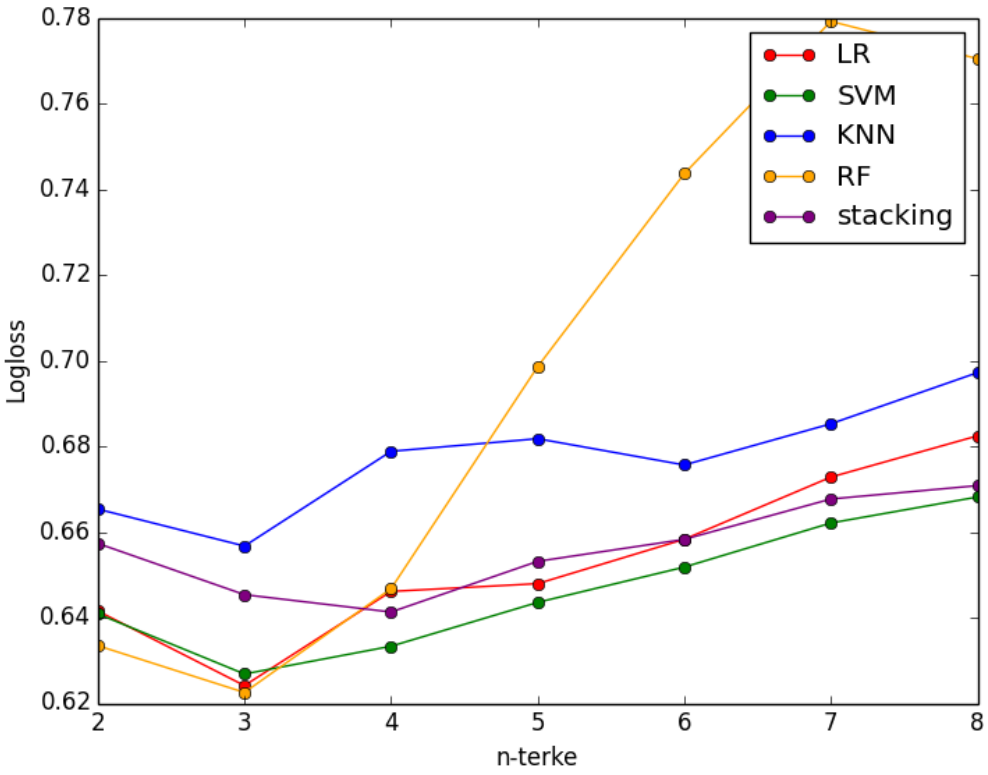
n-terka	2	3	4	5	6	7	8
LR	0.642	0.624	0.646	0.648	0.658	0.673	0.682
SVM	0.641	0.627	0.633	0.644	0.652	0.662	0.668
KNN	0.665	0.657	0.679	0.682	0.676	0.685	0.697
RF	0.634	**0.623**	0.647	0.699	**0.744**	0.779	0.770
Skladanje	0.657	0.645	0.641	0.653	0.658	0.668	0.671

Za boljši pregled nad tem, kako so se odrezale metode, predstavljamo še graf n-terk v odvisnosti od ocene logloss na sliki 5.1. Slika kaže, da se v večini najbolje obnese metoda podpornih vektorjev, najslabše pa metoda k najbljižjih sosedov.

Vse predstavljene ocene so bile pridobljene na podlagi ocene parametrov s tehniko interne validacije. Metoda naključnih gozdov je v 10-kratnem prečnem preverjanju največkrat pokazala najboljši rezultat pri 250-350 drevesih, logistična regresija se je najbolje obnesla pri $\lambda = 0.01$, metoda podpornih vektorjev pa pri $c = 200$.

Glede na oceno napovednih točnosti pa smo prišli do rezultatov, ki so prikazani v tabeli 5.2. Tudi tukaj pridemo do podobnih ugotovitev kot pri oceni logloss. Zmagovalna metoda naključnih gozdov se je najbolje obnesla pri parih črk z oceno 0.668, sledita pa ji logistična regresija s točnostjo 0.647 in metoda podpornih vektorjev s točnostjo 0.644, doseženi na trojkah črk.

Tudi tu opazimo, da za doseganje dovolj dobrih rezultatov niso primerni



Slika 5.1: Rezultati metod glede na oceno logloss pri napovedovanju razredov komentarjem.

podatki predstavljeni z največ črkami, v našem primeru z osmerkami. Najslabši rezultat je v tem primeru dosegla metoda k najbljižjih sosedov z oceno 0.594 pri osmerkah črk.

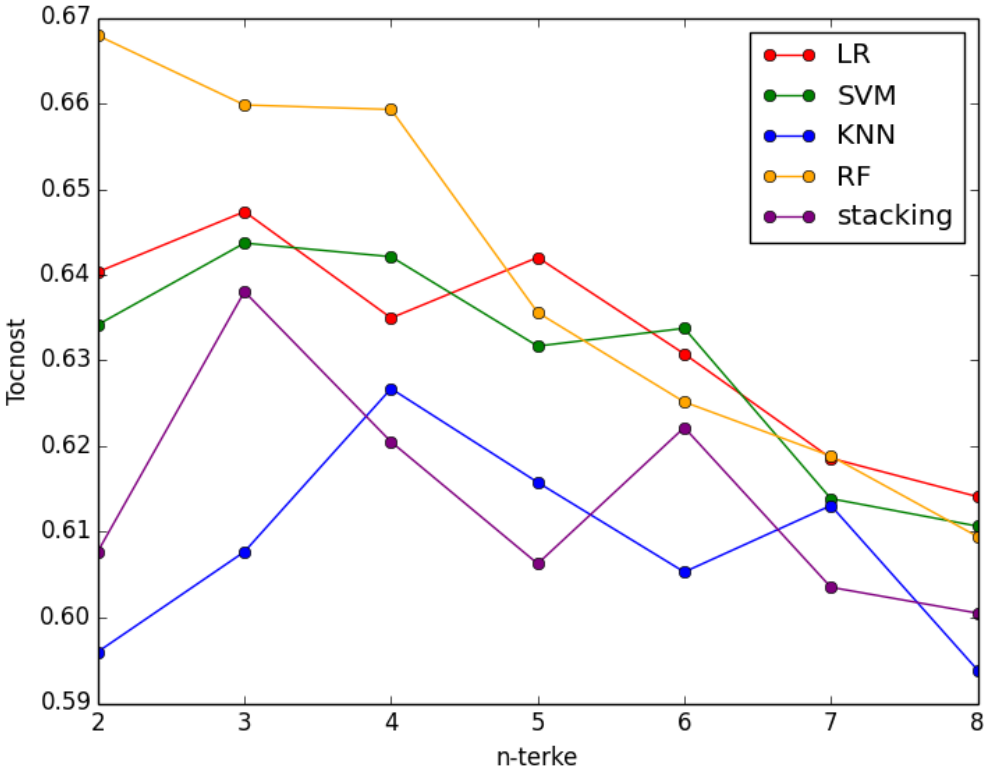
Še vedno velja, da je bila večina metod najbolje ocenjenih na podatkih, predstavljenih s trojkami črk, z izjemo metode k najbližjih sosedov, ki je najboljšo oceno dosegla pri podatkih s četvorkami črk, najslabše pa so bile metode ocenjene na podatkih iz osmerk črk.

Tabela 5.2: Rezultati metod glede na oceno točnosti pri razvrščanju komentarjev.

n-terka	2	3	4	5	6	7	8
LR	0.640	0.647	0.635	0.642	0.631	0.619	0.614
SVM	0.634	0.644	0.642	0.632	0.634	0.614	0.611
KNN	0.596	0.608	0.627	0.616	0.605	0.613	**0.594**
RF	**0.668**	0.660	0.659	0.636	0.625	0.619	0.610
Skladanje	0.608	0.638	0.621	0.606	0.622	0.604	0.601

Za boljšo predstavo si oglejmo še sliko 5.2, kjer je jasno razvidno, kako se metode obnašajo glede na različne vrste podatkov. Metoda naključnih gozdov tudi tu kaže največja odstopanja pri doseganju dobrih rezultatov.

Vrednosti parametrov, ki so botrovale k takšnim ocenam, so bile tudi v tem primeru skoraj enake, in sicer je metoda naključnih gozdov dosegla najboljši rezultat pri 350 drevesih, logistična regresija se je najbolje obnesla pri $\lambda = 0.01$, metoda podpornih vektorjev pa pri vrednosti parametra $c = 200$.



Slika 5.2: Rezultati metod glede na oceno točnosti pri napovedovanju razredov komentarjem.

5.2 Razprava

Numerične ocene točnosti napovednih modelov so lahko kazalec teže problema, ki ga modeliramo. Vsekakor je na mestu, da se vprašamo, zakaj so naše metode dosegale takšne rezultate in kaj to za nas pomeni.

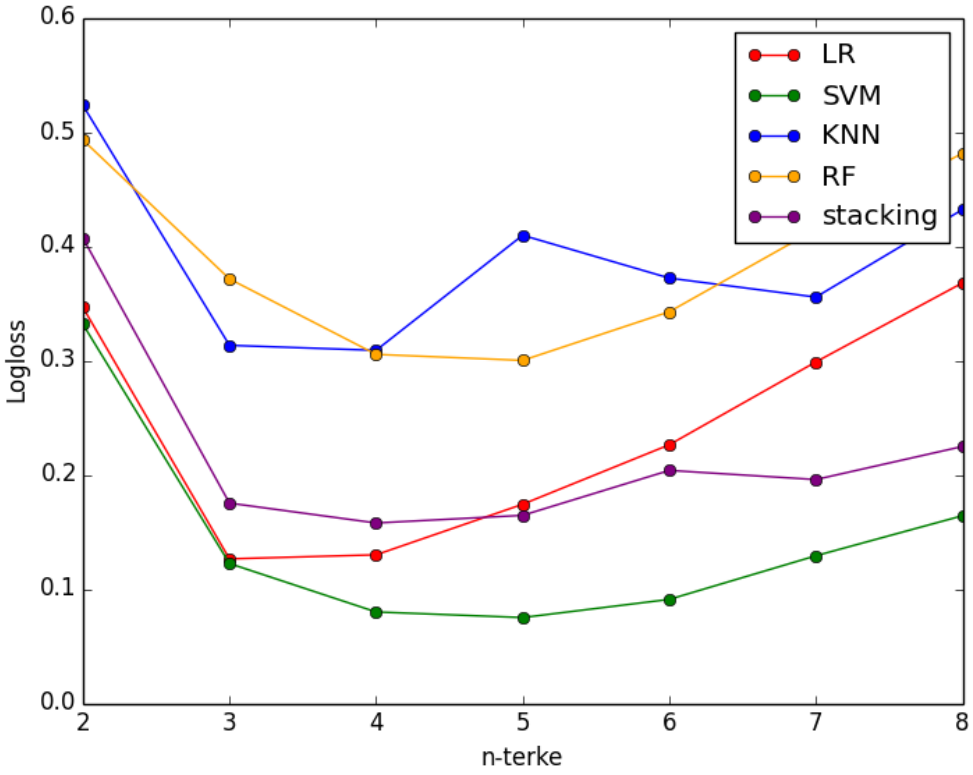
Pri klasifikacijskih problemih, kot je naš, je skoraj nemogoče zgraditi model, ki bo 100-odstotno natančen. V našem primeru lahko opazimo, da se delež pravilno napovedanih primerov giblje okoli vrednosti 65%. Naša pričakovanja so merila precej višje, zato bo potrebno preučiti še kaj, s čimer bomo lahko razložili zabeležene vrednosti ocen. Samo na podlagi teh rezultatov torej ne moremo govoriti, ali je to najboljše, kar lahko dosežemo pri raziskovanju naše problemske domene. Da bomo lahko podali konkretno oceno o tem, ali so te metode sploh primerne za reševanje tega problema in kaj jih je morebiti zmotilo pri doseganju boljših rezultatov, jih bomo preizkusili na podobnem problemu in te rezultate primerjali z že prej predstavljenimi.

Glede na oceno logloss pri preučevanju žanrov smo zabeležili uspešnosti napovednih modelov, ki so prikazane v tabeli 5.3. Najboljšo možno oceno je dosegla metoda podpornih vektorjev na podatkih, kjer so atribute predstavljale peterke črk, in sicer vrednost 0.075. Tudi pri drugih terkah je ista metoda dosegala precej boljše ocene kot druge metode. Za najslabši metodi sta se tokrat izkazali metodi k najbližjih sosedov in naključnih gozdov.

Skoraj vse metode so bile najbolje ocenjene na podatkih, predstavljenih s peterkami črk, najslabše pa na podatkih iz osmerk črk.

Poglejmo še grafično predstavitev na sliki 5.3. Razberemo lahko precej podoben trend pri vseh metodah, ki nam pove, da metode pokažejo najboljšo moč pri podatkih, predstavljenih s 4-5 črkami. Z atributi, ki jih predstavlja zelo majhno ali zelo veliko število črk, precej očitno izgubimo velik del informacije, ki bi nam pomagal pravilno razvrstiti primere.

Glede na oceno točnosti pri preučevanju žanrov smo zabeležili uspešnosti



Slika 5.3: Rezultati metod glede na oceno logloss pri napovedovanju razredov žanrom.

Tabela 5.3: Rezultati metod glede na oceno logloss pri razvrščanju žanrov.

n-terka	2	3	4	5	6	7	8
LR	0.347	0.126	0.130	0.174	0.226	0.299	0.368
SVM	0.332	0.122	0.080	**0.075**	0.091	0.129	0.164
KNN	0.523	0.313	0.310	0.410	0.372	0.356	0.432
RF	0.493	0.371	0.305	0.300	0.343	0.415	**0.481**
Skladanje	0.407	0.175	0.158	0.165	0.204	0.196	0.225

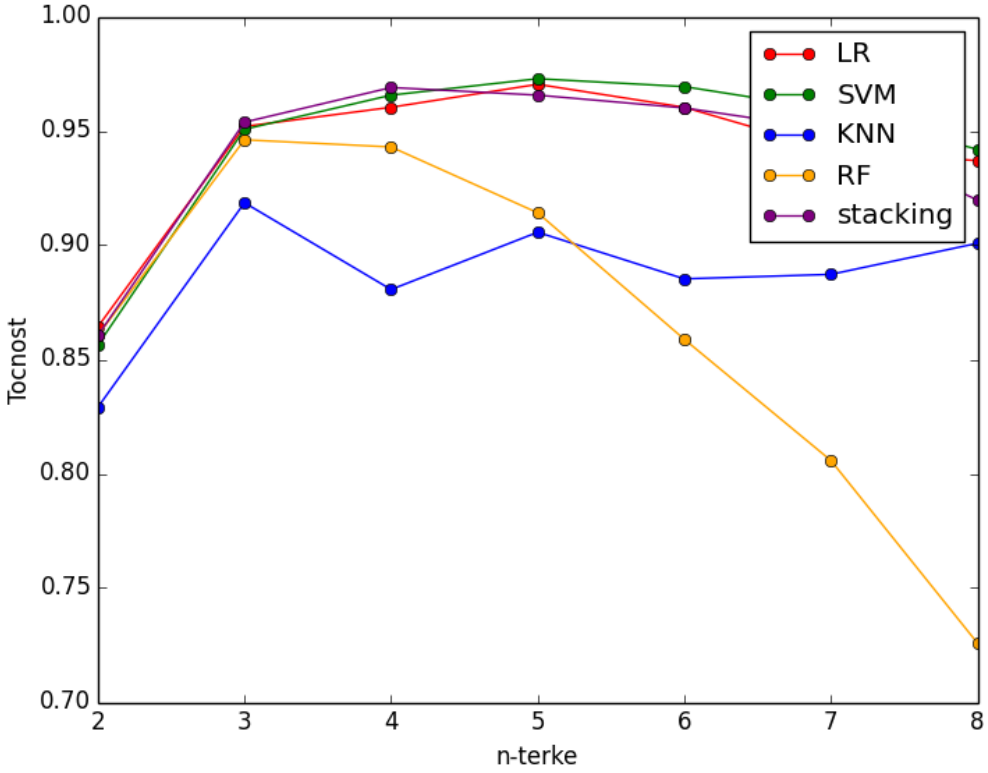
napovednih modelov, ki so prikazani v tabeli 5.4. Ponovno se izkaže, da je najboljša metoda podpornih vektorjev s točnostjo 0.973, najslabši pa sta metodi naključnih gozdov in k najbližjih sosedov.

Še vedno velja, da je bila večina metod najbolje ocenjenih na podatkih, predstavljenih s peterkami črk, opazimo pa tudi, da za doseganje dovolj dobrih rezultatov niso primerni podatki predstavljeni z najmanj in največ črkami, v našem primeru s pari in osmerkami. Najslabši rezultat je v tem primeru dosegla metoda naključnih gozdov z oceno 0.726.

Tabela 5.4: Rezultati metod glede na oceno točnosti pri razvrščanju žanrov.

n-terka	2	3	4	5	6	7	8
LR	0.864	0.952	0.960	0.970	0.960	0.942	0.937
SVM	0.856	0.951	0.966	**0.973**	0.969	0.959	0.942
KNN	0.829	0.919	0.881	0.906	0.885	0.887	0.901
RF	0.861	0.946	0.943	0.914	0.859	0.806	**0.726**
Skladanje	0.861	0.954	0.969	0.966	0.960	0.951	0.920

Za boljši vpogled preučimo še sliko 5.4. Vidimo lahko, da metoda podpornih vektorjev, logistična regresija in skladanje precej izstopajo, medtem ko se za veliko slabšo izkaže metoda k najbližjih sosedov, metoda naključnih gozdov pa spet prikaže ekstreme, s katerim odstopa od trenda, ki ga je moč



Slika 5.4: Rezultati metod glede na oceno točnosti pri napovedovanju razredov žanrom.

opaziti pri ostalih klasifikatorjih.

Če sedaj primerjamo rezultate na prvi in drugi problemski domeni, so razlike opazne. Pri žanrih so se vse metode odrezale mnogo bolje kot pri komentarjih. Lahko rečemo celo, da so v večini z več kot 95% pravilno napovedanimi primeri odlične tehnike za reševanje takih problemov. Potrdimo lahko torej hipotezo, da metode na komentarjih sicer dobro delujejo, vendar zaradi šlabih"podatkov ne dajejo pričakovanih rezultatov.

Prvo hipotezo smo torej potrdili, kar pomeni, da jedro naših težav ne leži v metodah, temveč v edini drugi možnosti - podatkih. Če logično razmislimo

o tem, zakaj je klasificiranje komentarjev glede na čustveno naravnanost njihovih avtorjev tako težko, je smiselno preučiti razlike med komentarji in članki, ki smo jih razvrščali v različna žanra. Kakšne težave po naši oceni lahko botrujejo h kvaliteti grajenja napovednih modelov, bomo predstavili v naslednjih odstavkih.

Najprej se lahko osredotočimo na samo rabo slovenskega jezika, kjer bomo že takoj opazili precejšnjo razliko. Avtorji komentarjev namreč ne uporabljajo knjižne slovenščine (gre bolj za zapise "po domače"oz. "pišejo kot govorijo"), zato o doslednosti uporabe jezika in slovnični pravilnosti le-tega ne moremo govoriti. Če pomislimo že na več kot 50 narečij, ki jih pozna slovenski jezik, lahko kaj kmalu ugotovimo, da gre za različno izražanje na več nivojih. Pri razvrščanju člankov v različne žanre pa lahko govorimo o visoki stopnji knjižne slovenščine (ki je ena in edina z razliko od prej omenjenih več deset narečij), saj tako avtorji člankov opravičujejo tudi kredibilnost napisanega. Sklepamo lahko, da imamo na eni strani torej neke nepravilnosti v podatkih, ki niso konsistentne in se nanašajo na nepravilno rabo slovenskega jezika, na drugi strani pa modele, ki iščejo podobnosti in sklepajo naprej na podlagi konsistence v podatkih. Ker se besede in samo izražanje v člankih uporabljajo bolj dosledno, je to lahko eden ključnih razlogov, zakaj modeli bolje klasificirajo besedila v žanre in ne glede na čustveno naravnanost.

Splošni problem, ki morda lahko nadaljuje razvoj zgornje teze in bi bil najbrž v takem smislu, kot mi predstavimo podatke (n-terke), zelo podoben prejšnjemu, je tudi to, da ima slovenščina sklanjatve, kar se seveda odraža pri različnemu tvorjenju besed. Čeprav gre v osnovi za eno samo osnovno besedo, katere koren ostaja enak, lahko variacije te besede na podlagi pripon in končnic pripišemo popolnoma drugim značilkam. V našem primeru torej atributi zavzamejo vse možne kombinacije teh besed in ne samo ene.

Iz vidika sintaktične pravilnosti sta to najbrž poglavitna razloga, zakaj prihaja do takšnih razlik, vendar pa je naš problem zelo verjetno težek zaradi

majhne semantične vrednosti, ki jo nosijo obravnavana besedila. V nadaljevanju bomo poskušali prikazati, da bistvo leži v pomenu samih besedil in osebni razlagi le-tega.

Ker je bilo razvrščanje komentarjev med pozitivne in negativne narejeno po subjektivni oceni, lahko obrazložimo razloge, zaradi katerih smo se tudi sami v določenih trenutkih znašli v dilemi, ali naj nek komentar pripišemo v pozitivni razred ali ne in obratno.

V precej komentarjih smo zasledili uporabo sarkazma, ki je seveda iz vidika matematičnih napovednih modelov precej problematična. Samo poved lahko celo napišemo tako, da iz vidika vsake posamezne besede lahko da čisto obraten vtis, kot če jo preberemo v kontekstu in ji na podlagi predznanja in razumevanja besedila kot celote pripišemo neko informacijo. Enako poved, lahko v enem primeru model klasificira kot pozitivno, v drugem pa kot negativno. Iz matematičnega vidika to pomeni 50-odstotno verjetnost za pripis primera v določen razred, kar je primerljivo s povsem naključnim razvrščanjem. Z razliko od človeškega razuma matematične metode ne morejo zaznati tona, v katerem je bila izjava podana, saj za negativen prizvok niti niso potrebne točno določene besede, ki jih smatramo kot negativne. Prav zaradi tega lahko izbrane metode ne dajejo pričakovanih rezultatov.

Ročno razvrščanje komentarjev pa je bilo problematično tudi iz vidika, kako komentatorji razumejo bistvo članka. V komentarjih se pogosto razvijejo debate med komentatorji, katerih tema ni nujno to, o čemer govori članek, vendar kaj sorodnega, s čimer želijo komentatorji opozoriti na podobnost drugih tem oz. problematik. V komentarjih smo zasledili tudi pogosto spuščanje na osebno raven med dvema ali več komentatorji, kar se je na koncu odražalo v popolni zgrešitvi teme, ki naj bi bila jedro članka. Taka besedila so že za nas predstavljala problem, pri matematičnem obravnavanju le-tega pa padejo povsem ven iz konteksta in niso relevantna za problem, ki ga raziskujemo.

Če gledamo s stališča žanrov, je najbrž povsem razumljivo, da pri določenih temah obstajajo besede, ki so značilne za določen žanr, zato je izbira, v kateri razred spada neko besedilo, precej očitna in posledično tudi lažja, kot pri ocenjevanju tega, ali je neko osebno mnenje izrazito pozitivno ali negativno. Mnogokrat se v komentarjih pojavijo deljena mnenja, ki nekatere vidike pohvalijo, spet druge pa grajajo, zato je stopnja težavnosti tega problema še toliko večja. Temu bi se lahko izognili z razvrščanjem primerov v nevtralni razred na način, ki ga opisuje Koppel [10], vendar naša problemska domena s približno 500 primeri ne bi mogla zagotoviti zadostno število resnično pozitivnih in negativnih primerov.

5.3 Statistična primerjava klasifikatorjev

V prejšnjem poglavju smo se osredotočali le na en nabor podatkov - dotično nterko, in želeli ugotoviti, katera metoda je najbolj primerna za katere podatke ter s katero lahko dosežemo najvišjo stopnjo pravilnega napovedovanja. Naš cilj pa je, da na koncu poročamo, katera metoda ali več njih se v splošnem najbolje obnesejo.

Za statistično analizo smo izbrali postopek, ki ga v svojem delu opiše Demšar [8] in je primeren za primerjavo več klasifikatorjev na več naborih podatkov. V prvem delu bomo za ovrednotenje uporabili neparametrični Friedmanov test, s katerim bomo potrdili ali zavrgli ničelno hipotezo. Za bolj natančno nadaljnjo analizo bo služil Nemenyijev test, na koncu pa bomo ugotovitve prikazali še z grafom kritične razdalje, ki smo ga izrisali s pomočjo programskega sistema Orange[1].

Friedmanov test rangira metode za vsak nabor podatkov posamezno - z oceno od 1 do k ocenimo, kako so se metode odrezale pri posameznem naboru

[1] `http://orange.biolab.si/docs/latest/reference/rst/Orange.evaluation.scoring/`

podatkov, kjer 1 pomeni, da se je metoda odrezala najbolje, k pa najslabše. Nato test primerja povprečne range metod z ničelno hipotezo, ki pravi, da so vse metode enako dobre. Friedmanova statistiko izračunamo po enačbi:

$$\chi_F^2 = \frac{12N}{k\,(k+1)} \left[\sum_j R_j^2 - \frac{k\,(k+1)^2}{4} \right] \tag{5.1}$$

kjer N pomeni število podatkovnih naborov, k število klasifikatorjev, R_j pa povprečni rang metode na podatkih.

Ker pa je bilo ugotovljeno, da je ta statistika precej konzervativna, bomo pri izračunu uporabili še izboljšavo le- te, ki je prikazana z enačbo:

$$F_F = \frac{(N-1)\,\chi_F^2}{N\,(k-1) - \chi_F^2} \tag{5.2}$$

in je porazdeljena glede na F porazdelitev s $k-1$ in $(k-1)(N-1)$ stopnjama prostosti. Tabela kritičnih vrednosti je splošno znana.

Če je bila ničelna hipoteza na zgoraj opisani način zavrnjena, lahko nadaljujemo z nadaljnjimi testi. Nemenyijev test se uporablja ravno pri primerjanju več klasifikatorjev med sabo.

Uspešnost dveh klasifikatorjev je bistveno drugačna, če se pripadajoča povprečna ranga med seboj razlikujeta vsaj za kritično razdaljo

$$CD = q_\alpha \sqrt{\frac{k\,(k+1)}{6N}} \tag{5.3}$$

kjer kritične vrednosti q_α lahko razberemo iz porazdelitve t-testa in jih prilagodimo tako, da jih delimo s $\sqrt{2}$. Prilagojene vrednosti so prikazane v tabeli 5.5.

V naslednjih podpoglavjih bomo predstavili podrobno analizo uspešnosti klasifikatorjev na komentrajih in žanrih.

5.3.1 Komentarji

V tem poglavju bomo analizirali, kako so se metode odrezale na podatkih, pridobljenih iz komentarjev, najprej glede na oceno logloss in nato še glede

q_α	Št. klasifikatorjev = 5
$q_{0.05}$	2.728
$q_{0.10}$	2.459

Tabela 5.5: Kritične vrednosti za test Nemenyi za 5 klasifikatorjev

na točnost.

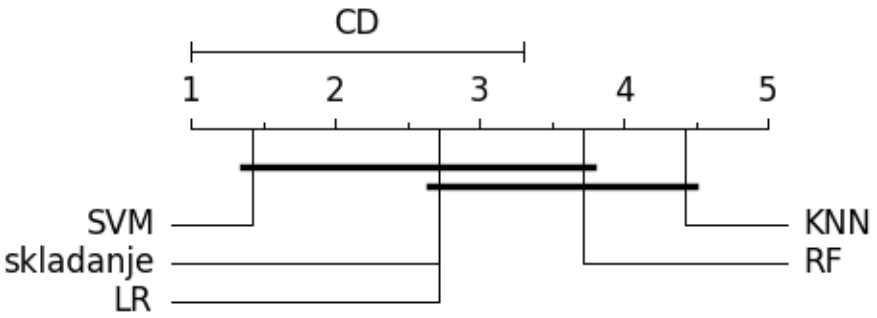
Najprej vse rezultate zberemo v tabeli in jih za vsak posamezen set podatkov rangiramo glede na to, katera metoda se je izkazala najbolje in katera najslabše. Postopek je prikazan v tabeli 5.6.

Tabela 5.6: Prikaz rangiranja metod pri Friedmanovem testu na podlagi ocene logloss pri razvrščanju komentarjev.

n-terke	LR		SVM		KNN		RF		Skladanje	
2	0,642	3	0,641	2	0,665	5	0,633	1	0,657	4
3	0,624	2	0,627	3	0,657	5	0,623	1	0,645	4
4	0,646	3	0,633	1	0,679	5	0,647	4	0,641	2
5	0,648	2	0,644	1	0,682	4	0,699	5	0,653	3
6	0,658	3	0,652	1	0,676	4	0,744	5	0,658	2
7	0,673	3	0,662	1	0,685	4	0,779	5	0,668	2
8	0,682	3	0,668	1	0,697	4	0,770	5	0,671	2
Povprečni rang	2,714		1,429		4,429		3,714		2,714	

Friedmanov test preveri ali se povprečni rangi bistveno razlikujejo od povprečnega ranga $R_j = 3$, ki je določen z ničelno hipotezo:

$$\chi_F^2 = \frac{12 \cdot 7}{5\,(5+1)} \left[\left(2.714^2 + 1.429^2 + 4.429^2 + 3.714^2 + 2.714^2\right) - \frac{5\,(5+1)^2}{4} \right] = 14.51 \tag{5.4}$$



Slika 5.5: Graf kritične razdalje glede na oceno logloss pri završčanju komentarjev za $\alpha = 0.05$.

$$F_F = \frac{(7-1) \cdot 14.51}{7\,(5-1) - 14.51} = 6.45 \tag{5.5}$$

S petimi metodami in sedmimi podatkovnimi nabori je F_F vrednost porazdeljena s F porazdelitvijo s $5-1=4$ in $(5-1) \times (7-1) = 24$ stopnjama prostosti. Kritična vrednost za $F(4, 24)$ za $\alpha = 0.05$ je 2.31, zato lahko ničelno hipotezo zavrnemo.

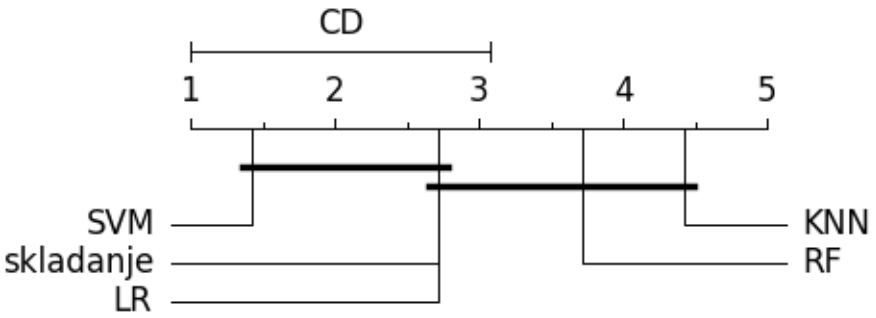
Nadaljno analizo nato izvedemo z Nemenyijevim testom. Kritična vrednost za $\alpha = 0.05$ je 2.728 (tabela 5.5). Izračunamo kritično razdaljo

$$CD = 2.728\sqrt{\frac{5\,(5+1)}{6 \cdot 7}} = 2.31 \tag{5.6}$$

Ker je kritična razdalja krajša od razdalje med najboljšo in najslabšo metodo, bo ta primerjava zadostna za bistveno razlikovanje med algoritmi.

Rezultate nato predstavimo z grafom kritične razdalje na sliki 5.5, iz katerega bomo lažje nazorno predstavili naše ugotovitve.

Glede na izračunane razdalje med algoritmi in graf kritične razdalje lahko sklepamo, da se metoda podpornih vektorjev bistveno razlikuje od metode najbližjih sosedov, pri čemer je prva bistveno boljša od druge. Za skladanje, logistično regresijo in metodo naključnih gozdov pa na podlagi naših



Slika 5.6: Graf kritične razdalje glede na oceno logloss pri završčanju komentarjev za $\alpha = 0.10$.

rezultatov ne moremo trditi, da se med sabo značilno razlikujejo.

Če vrednost α povečamo na 0.10 in s tem zajamemo večji vzorec, se izkaže, da je metoda podpornih vektorjev bistveno boljša od metode najbližjih sosedov in naključnih gozdov, za skladanje in logistično regresijo pa še vedno ne moremo govoriti o bistvenih razlikah. Razlike so prikazane na sliki 5.6.

Pri rezultatih metod glede na oceno točnosti postopek ponovimo. Izračunani rangi so prikazani v tabeli 5.7.

$$\chi_F^2 = \frac{12 \cdot 7}{5\,(5+1)} \left[\left(1.857^2 + 2.429^2 + 4.571^2 + 1.714^2 + 4.429^2\right) - \frac{5\,(5+1)^2}{4} \right] = 21.83 \tag{5.7}$$

$$F_F = \frac{(7-1) \cdot 21.83}{7\,(5-1) - 21.83} = 21.23 \tag{5.8}$$

Kritična vrednost F porazdelitve ostaja enaka, prav tako pa tudi kritična razdalja. Ker je slednja tudi v tem primeru krajša od razdalje med najboljšo in najslabšo metodo, bo ta primerjava zadostna za bistveno razlikovanje med algoritmi.

Rezultati so predstavljeni z grafom kritične razdalje na sliki 5.7.

Tabela 5.7: Prikaz rangiranja metod pri Friedmanovem testu na podlagi ocene točnosti pri razvrščanju komentarjev.

n-terka	LR		SVM		KNN		RF		Skladanje	
2	0,640	2	0,634	3	0,596	5	0,668	1	0,608	4
3	0,647	2	0,644	3	0,608	5	0,660	1	0,638	4
4	0,635	3	0,642	2	0,627	4	0,659	1	0,621	5
5	0,642	1	0,632	3	0,616	4	0,636	2	0,606	5
6	0,631	2	0,634	1	0,605	5	0,625	3	0,622	4
7	0,619	2	0,614	3	0,613	4	0,619	1	0,604	5
8	0,614	1	0,611	2	0,594	5	0,609	3	0,601	4
Povprečni rang	1,857		2,429		4,571		1,714		4,429	

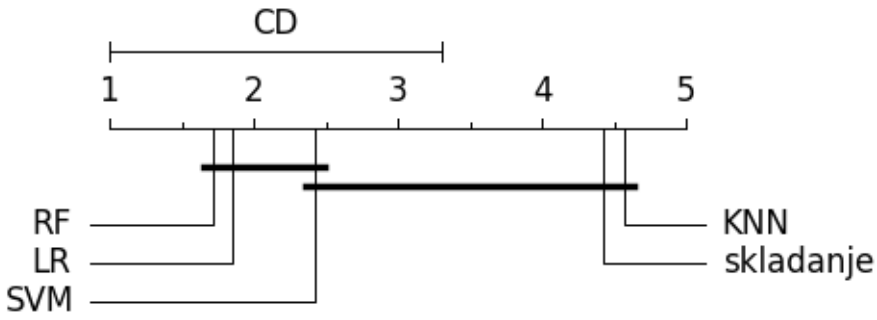
Glede na oceno točnosti lahko povzamemo, da sta metoda naključnih gozdov in logistična regresija značilno boljši od metode najbližjih sosedov in skladanja. Za metodo podpornih vektorjev pa v tem trenutku eksperimentalni podatki ne zadostujejo za podajanje kakršne koli trditve o bistvenem razlikovanju.

Ob povečanju vrednosti α na 0.10 na sliki 5.8 ne opazimo sprememb.

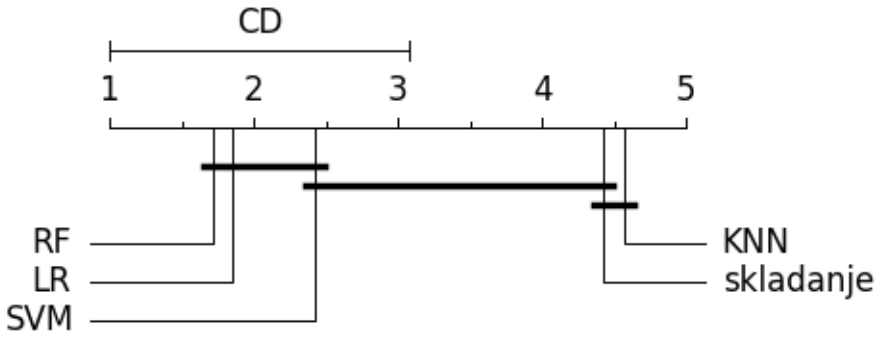
Če sedaj primerjamo metode glede na obe predstavljeni oceni, lahko v splošnem povzamemo, da bo za naš problem od teh petih algoritmov metoda najbližjih sosedov vedno najslabša izbira. Metoda podpornih vektorjev in logistična regresija pa spadata v skupino boljših metod.

5.3.2 Žanri

V tem poglavju bomo analizirali, kako so se metode odrezale na podpornih podatkih, pridobljenih iz žanrov, najprej glede na oceno logloss in nato še glede na točnost. Rezultati rangiranja so prikazani v tabeli 5.8.



Slika 5.7: Graf kritične razdalje glede na oceno točnosti pri završčanju komentarjev za $\alpha = 0.05$.



Slika 5.8: Graf kritične razdalje glede na oceno točnosti pri završčanju komentarjev za $\alpha = 0.10$.

Tabela 5.8: Prikaz rangiranja metod pri Friedmanovem testu na podlagi ocene logloss pri razvrščanju žanrov.

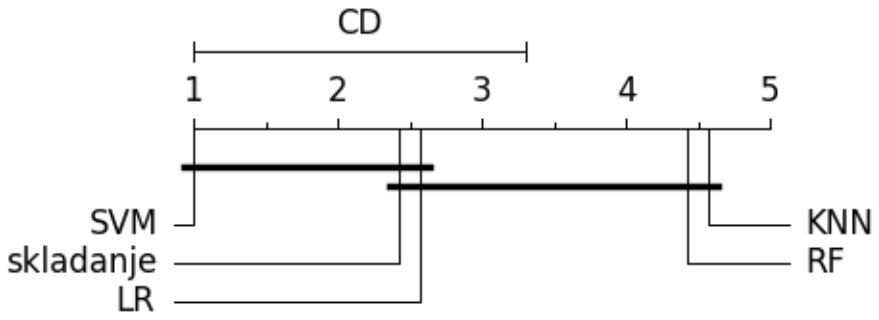
n-terka	LR		SVM		KNN		RF		Skladanje	
2	0,347	2	0,332	1	0,523	5	0,493	4	0,407	3
3	0,126	2	0,122	1	0,313	4	0,371	5	0,175	3
4	0,130	2	0,080	1	0,309	5	0,305	4	0,158	3
5	0,174	3	0,075	1	0,410	5	0,300	4	0,164	2
6	0,226	3	0,091	1	0,372	5	0,343	4	0,204	2
7	0,299	3	0,129	1	0,355	4	0,415	5	0,196	2
8	0,368	3	0,164	1	0,432	4	0,481	5	0,225	2
Povprečni rang	2,571		1		4,571		4,429		2,429	

$$\chi_F^2 = \frac{12 \cdot 7}{5\,(5+1)} \left[\left(2.571^2 + 1^2 + 4.571^2 + 4.429^2 + 2.429^2\right) - \frac{5\,(5+1)^2}{4} \right] = 25.26 \tag{5.9}$$

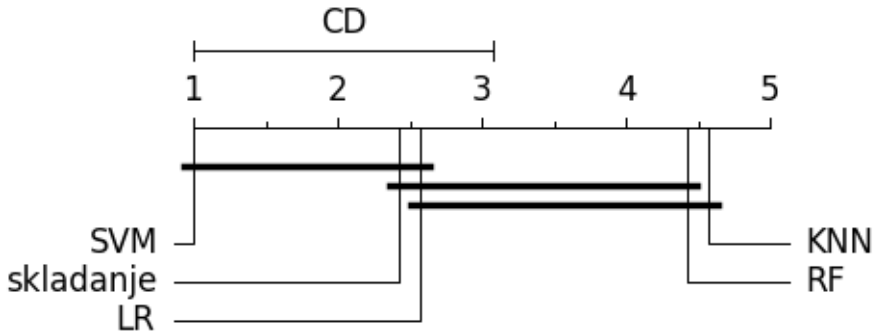
$$F_F = \frac{(7-1) \cdot 25.26}{7\,(5-1) - 25.26} = 55.31 \tag{5.10}$$

Kritična vrednost F porazdelitve še vedno ostaja enaka, zato lahko tudi v tem primeru zavrnemo ničelno hipotezo. Prav tako kritična razdalja za $\alpha = 0.05$ zadostuje za bistveno razlikovanje med algoritmi. Graf kritične razdalje na sliki 5.9 podrobno prikaže razlikovanja med metodami.

Zaključimo lahko, da se metoda podpornih vektorjev po napovedni točnosti značilno razlikuje od metode najbližjih sosedov in naključnih gozdov, pri čemer je prva bistveno boljša od drugih dveh. Za skladanje in logistično regresijo pa v tem trenutku eksperimentalni podatki ne zadostujejo za podajanje kakršne koli trditve o bistvenem razlikovanju.



Slika 5.9: Graf kritične razdalje glede na oceno logloss pri završčanju žanrov za $\alpha = 0.05$.



Slika 5.10: Graf kritične razdalje glede na oceno logloss pri završčanju žanrov za $\alpha = 0.10$.

Če vrednost α povečamo na 0.10, lahko dodamo še ugotovitev, da je metoda najbližjih sosedov bistveno slabša od metode podpornih vektorjev in skladanje. Razlike so prikazane na sliki 5.10.

Pri rezultatih metod glede na oceno točnosti postopek ponovimo. Izračunani rangi so prikazani v tabeli 5.9.

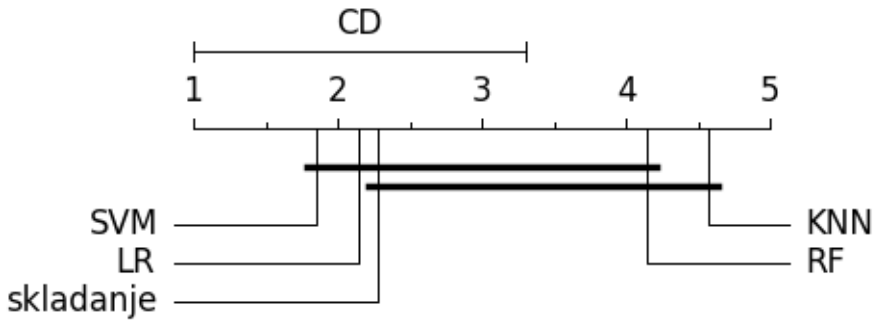
Tabela 5.9: Prikaz rangiranja metod pri Friedmanovem testu na podlagi ocene točnosti pri razvrščanju žanrov.

n-terka	LR		SVM		KNN		RF		Skladanje	
2	0,864	1	0,856	4	0,829	5	0,861	2	0,861	3
3	0,952	2	0,951	3	0,919	5	0,946	4	0,954	1
4	0,960	3	0,966	2	0,881	5	0,943	4	0,969	1
5	0,970	2	0,973	1	0,906	5	0,914	4	0,966	3
6	0,960	2	0,969	1	0,885	4	0,859	5	0,960	3
7	0,942	3	0,959	1	0,887	4	0,806	5	0,951	2
8	0,937	2	0,942	1	0,901	4	0,726	5	0,920	3
Povprečni rang	2,143		1,857		4,571		4,143		2,286	

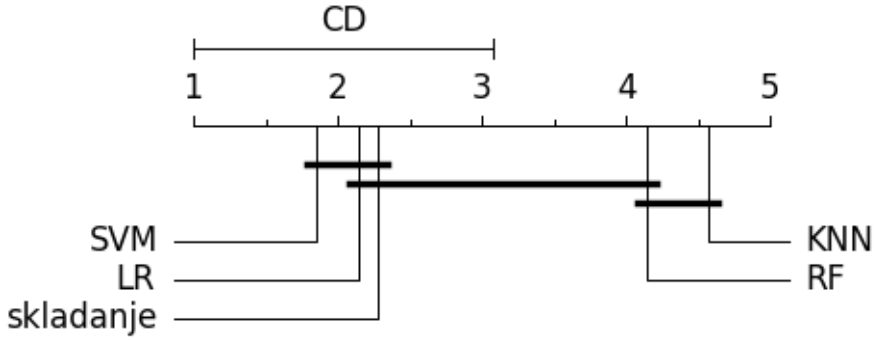
$$\chi_F^2 = \frac{12 \cdot 7}{5\,(5+1)} \left[\left(2.143^2 + 1.857^2 + 4.571^2 + 4.143^2 + 2.286^2\right) - \frac{5\,(5+1)^2}{4} \right] = 17.71 \tag{5.11}$$

$$F_F = \frac{(7-1) \cdot 17.71}{7\,(5-1) - 17.71} = 10.33 \tag{5.12}$$

Kritična vrednost F porazdelitve ostaja enaka, prav tako pa tudi kritična razdalja. Ker je slednja tudi v tem primeru krajša od razdalje med najboljšo in najslabšo metodo, bo ta primerjava zadostna za bistveno razlikovanje med algoritmi.



Slika 5.11: Graf kritične razdalje glede na oceno točnosti pri završčanju žanrov za $\alpha = 0.05$.



Slika 5.12: Graf kritične razdalje glede na oceno točnosti pri završčanju žanrov za $\alpha = 0.10$.

Rezultati so predstavljeni z grafom kritične razdalje na sliki 5.11.

Glede na oceno točnosti lahko povzamemo, da sta metoda podpornih vektorjev in logistična regresija bistveno boljši od metode najbližjih sosedov in metode skladanje. Za skladanje in naključne gozdove pa v tem trenutku eksperimentalni podatki ne zadostujejo za podajanje kakršne koli trditve o bistvenem razlikovanju.

Ob povečanju vrednosti α na 0.10 na sliki 5.12 se pokaže tudi bistvena razlika med uspešnostjo metode najbližjih sosedov in skladanja.

Pri primerjavi metod glede na obe oceni, lahko rečemo, da metodi najbližjih sosedov in naključni gozdovi nista primerni za takšno problemsko domeno. Metoda podpornih vektorjev je absolutni zmagovalec, daleč za njo pa ne zaostajata tudi skladanje in logistična regresija.

5.3.3 Primerjava metod na podlagi obeh podatkovnih domen

Iz prejšnjih izračunov kritične razdalje, smo za vsako domeno posebej lahko precej natančno ovrednotili, katere metode se obnesejo bolje in katere za raziskovanje takih problemov niso najboljša izbira. Če pa želimo na splošno povedati, katere metode so za take probleme najbolj primerne, moramo poiskati skupne točke prve in druge. Opazili smo lahko, da se metoda naključnih gozdov pri komentarjih izkaže kot zelo dobra, pri žanrih pa kot slabša klasifikacijska metoda, zato ne moremo zagotovo govoriti, ali je na mestu za podajanje kakršnih koli ugotovitev. Za metodo k najbližjih sosedov lahko z gotovostjo trdimo, da se vedno obnese precej slabše kot ostale metode, metoda podpornih vektorjev pa je vedno v boju za prvo mesto med klasifikatorji. Logistična regresija in skladanje ne zaostajata veliko, vendar pa skladanje v splošnem tu ne pripomore k izboljšanju rezultatov.

Poglavje 6

Sklepne ugotovitve

V zadnjem času ljudje čedalje bolj izražamo svoje mnenje o kakršni koli temi predvsem na svetovnem spletu, saj tu nismo tako osebno izpostavljeni. Internet je tako postal zbiralnica različnih besedil, iz katerih se je možno marsičesa naučiti. Raziskovanja na tem področju vodijo v smer avtomatičnega razvrščanja takšnih tekstov. Začelo se je s klasifikacijo besedil v različne žanre, kjer želimo na podlagi značilnih besed napovedati, o kateri temi besedilo govori (npr. ali gre za vsebino politične ali športne narave). Ta razlikovanja so precej očitna in tudi za človeka precej enostavna. Zadnji trendi pa gredo v smer, kako in na podlagi česa bi lahko s pomočjo računalnika in primernih algoritmov znali avtomatično zaznati, kakšno čustveno stanje izraža avtor v besedilu - ali kaže naklonjenost temi, o kateri govori, ali se morda z njo ne strinja.

V diplomskem delu smo ugotovili, da so za klasifikacijo pozitivnih in negativnih mnenj primerne predvsem metode, ki tudi pri razvrščanju v različne teme dajejo dobre rezultate. Kljub temu, da te metode znajo po principu razvrščanja glede na teme primere v več kot 90% točno razvrstiti, pa se pri ugotavljanju pripadnosti med negativne ali pozitivne pokaže precejšnje odstopanje. Izbrane tehnike strojnega učenja so v le dobrih 60% znale napovedati, ali se primer uvršča med pozitivne ali negativne, glede na to, da je zastopa-

nost razredov predstavlja razmerje 6:4 za negativni razred. Med razloge za tako velike razlike so se uvrstili problemi sintaktične in semantične narave, ki smo jih opisali pri vrednotenju rezultatov. Ključna stvar pri sintaksi je ta, da pri komentarjih ne moremo govoriti o dosledni uporabi slovenščine, pri semantiki pa to, da je že za človeški razum včasih zaznava tona v besedilu težka, kar pa pomeni še toliko večji problem za matematični klasifikacijski model. V diplomski nalogi smo pokazali, da se pri tovrstnih primerih od obravnavanih metod najbolje obneseta metoda podpornih vektorjev in logistična regresija. Precej dobre rezultate lahko poda tudi metoda skladanja, ki z lahkoto prekaša drugo uvrščeno metodo, žal pa v našem primeru ni izboljšala rezultatov že najboljše metode.

Če primerjamo način klasifikacije v razrede tega diplomskega dela z drugo že prej omenjeno diplomsko nalogo [3], je razlik kar nekaj. Pri obdelavi komentarjev je v obeh delih uporabljen podoben pristop, res pa se predhodnje delo poleg strojnega učenja poslužuje še prej pripravljenega korpusa, ki besede zamenja z njihovimi lemami, nato pa te dokumente predstavi z vrečo besed in le-to pretvori s transformacijo TF-IDF. V tem delu z različnimi klasifikacijskimi tehnikami pridejo do nekoliko, najbrž neznačilno boljših rezultatov, ki smo jih prikazali tudi mi s predstavitvami podatkov v obliki n-terk. Prav tako se v obeh delih za najboljšo metodo v večini primerov izkaže metoda podpornih vektorjev, metoda k najbližjih sosedov pa za najslabšo. Do večjih razlik v rezultatih pride kasneje z upoštevanjem nevtralnega razreda, ki se izkaže za pomemben del te raziskave, a smo ga v našem delu zaradi lažjega razumevanja zanemarili. Metoda podpornih vektorjev je v tem primeru dosegla precej višjo klasifikacijsko točnost, in sicer 82 %. Do boljših rezultatov so v drugem delu prišli tudi na podlagi zajema veliko več podatkov in tematike z bolj enakomerno porazdelitvijo razredov. V obzir jemljejo več člankov in njihovih komentarjev, mi pa analize izvedemo le na podlagi komentarjev enega političnega članka, kjer se izkaže, da prevladujejo predvsem

negativni komentarji.

Problem, ki smo ga raziskovali, je predmet čedalje večje obravnave. Podoben primer je bilo moč zaslediti npr. v času zadnjih predsedniških volitev, kjer so mnenja o posameznih kandidatih s strani ljudi, ki so o tem razglabljali na socilnih omrežjih, zajemali in na njih izvajali različne klasifikacijske algoritme (Gama System® PerceptionAnalytics[1]). S pomočjo teh metod so lahko sproti avtomatično napovedovali, kakšen je trend, kateri posameznik ima pri volivcih največjo podporo in kako se je ta na podlagi različnih dogodkov in soočenj kandidatov v času volilnih kampanij tudi spreminjala. Rezultati analiz so bili objavljeni na spletnem mestu `http://predsedniskevolitve.si/`.

To je le eden od bolj znanih primerov, ki so se pojavili tudi na slovenskem območju, sicer pa so takšne statistične obravnave besedil lahko pomembne tudi v poslovnem svetu. Danes podjetja na trg prihajajo z raznoraznimi produkti in storitvami, vsak tak prodor pa je lahko uspešen ali neuspešen glede na to, kako je sprejet v svetu potrošnikov. Seveda je cilj vsake novosti na trgu ponuditi rešitev za potrošnikov problem, zaradi česar bo le-ta kupil proizvod ali storitev, podjetju pa s tem prinesel dobiček. Trg je dandanes zelo nepredvidljiv in mnenje skupine posameznikov v določen podjetju ne odraža nujno mnenja ljudi, ki bi določen produkt tudi kupili. Cilj je torej optimizirati poslovni proces in na tržišče poslati produkt, ki ima večjo možnost za uspeh, in odstraniti izdelke, ki podjetju prinašajo le izgubo. Neko podjetje bi tako pred veliko izdajo novega produkta lahko na podlagi takšne raziskave ugotovilo, ali bo le-ta dobro sprejet med potrošniki, in temu primerno prilagodilo samo strategijo vloženih sredstev in prodaje izdelka. Podobno raziskavo o mnenju ljudi iz določenih aspektov izdelkov so opisali tudi Lopezova in drugi [1] ter s tem pokazali pomembnost raziskovanja mnenj v poslovnem svetu.

Ker smo ugotovili, da je problem takšne narave precej trd oreh, ima raziskovanje v tej smeri še precej odprtih poti. Če se osredotočimo na reševanje

[1] `http://www.gama-system.si/sl/produkti/gama-system-perceptionanalytics`

dotičnega problema v tej diplomski nalogi, bi bilo gotovo vredno raziskati posledice drugačne priprave podatkov. Za začetek bi komentarje morda lahko napisali v knjižni slovenščini in opazovali, kako veliko vlogo ima pri taki klasifikaciji res sintaksa in kako velika teža temelji na semantiki. Drugi korak pa bi bil lahko splošno klasificiranje posameznih besed med take s pozitivnim ali take z negativnim prizvokom s predhodno obdelavo, kjer bi vse variacije besede predstavljale eno entiteto.

Literatura

[1] Alejandra Lopez; Tony Veale and Prasenjit Majumder, "Feature Extraction from Product Reviews using Feature Similarity and Polarity", *Technical Report UCD-CSI-2009-09*, 2009.

[2] Bo Pang; Lillian Lee and Shivakumar Vaithyanathan, "Thumbs up? Sentiment Classification using Machine Learning Techniques", *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP).*, str. 79—86, 2002.

[3] Brina Škoda, "Rudarjenje razpoloženja na komentarjih rtvslo.si", *Diplomsko delo univerzitetnega študija. Univerza v Ljubljani, Fakulteta za računalništvo in informatiko*, 2013.

[4] Peng, Chao-Ying Joanne; Lee, Kuk Lida; Ingersoll and Gary M., "An Introduction to Logistic Regression Analysis and Reporting.", *Journal of Educational Research, v96 n1*, str. 3–14, 2002.

[5] D.Karthika Renuka Dr.T.Hamsapriya, "Email classification for Spam Detection using Word Stemming", *International Journal of Computer Applications 1(5)*, str. 45—47, 2010.

[6] David H. Wolpert, "Stacked generalization", *Neural Networks, 5(2)*, str. 241–259, 1992.

[7] Getoor, Lise and Segal, Eran and Taskar, Benjamin and Koller, Daphne, "Probabilistic Models of Text and Link Structure for Hypertext Classification", *IJCAI Workshop on Text Learning: Beyond Supervision*, 2001.

[8] Janez Demšar, "Statistical Comparisons of Classifiers over Multiple Data Sets", *Journal of Machine Learning Research 7*, str. 1—30, 2006.

[9] L. Breiman, "Random Forests", *Machine Learning, 45(1)*, str. 5—32, 2001.

[10] Moshe Koppel; Jonathan Schler, "The Importance of Neutral Examples for Learning Sentiment", *Computational Intelligence 22*, str. 100—109, 2006.

[11] Paulo Gaspar; Jaime Carbonell and Jose Luis Oliveira, "On the parameter optimization of Support Vector Machines for binary classification", *Journal of Integrative Bioinformatics, 9(3):201*, 2012.

[12] Peter Turney, "Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews", *Proceedings of the Association for Computational Linguistics.*, str. 417—424, 2002.

[13] Rada Mihalcea; Carmen Banea and Janyce Wiebe, "Learning Multilingual Subjective Language via Cross-Lingual Projections", *Proceedings of the Association for Computational Linguistics (ACL).*, str. 976—983, 2007.

[14] Soo-Min Kim; Eduard Hovy, "Automatic identifcation of pro and con reasons in online reviews", *Proceedings of the Association for Computational Linguistics (ACL).*, str. 483—490, 2006.

[15] Vinko Vodopivec, "Statistična primerjava črk in besed", *Jezikovni temelji starejše slovenske etnogeneze, Jutro, Ljubljana*, str. 16-30, 2010.

[16] Xiaogang Han; Junfa Liu, Zhiqi Shen and Chunyan Miao, "An Optimized K-Nearest Neighbor Algorithm for Large Scale Hierarchical Text Classification", *Joint ECML/PKDD PASCAL Workshop on Large-Scale Hierarchical Classication.*, str. 2–12, 2011.

Dodatek A

Logistična regresija - rezultati

Tabela A.1: Napovedne točnosti na podlagi ocene logloss za logistično regresijo, izvedeno na komentarjih.

	n-terka	2	3	4	5	6	7	8
	ZP	**0.01**	**0.01**	**0.001**	**0.001**	**0.001**	**0.001**	**0.01**
1	OP	0.634	0.629	0.633	0.645	0.661	0.671	0.678
	OUM	0.692	0.643	0.649	0.631	0.628	0.639	0.67
	ZP	**0.01**	**0.01**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**
2	OP	0.663	0.643	0.651	0.651	0.663	0.670	0.675
	OUM	0.577	0.567	0.556	0.576	0.608	0.638	0.655
	ZP	**0.1**	**0.01**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**
3	OP	0.665	0.632	0.612	0.629	0.649	0.661	0.670
	OUM	0.634	0.612	0.646	0.645	0.653	0.662	0.675
	ZP	**0.01**	**0.01**	**0.01**	**0.001**	**0.001**	**0.001**	**0.001**
4	OP	0.659	0.645	0.653	0.656	0.661	0.668	0.675
	OUM	0.575	0.574	0.616	0.603	0.625	0.631	0.636
	ZP	**0.01**	**0.01**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**
5	OP	0.643	0.634	0.632	0.644	0.657	0.664	0.670
	OUM	0.651	0.618	0.655	0.659	0.662	0.681	0.691
	ZP	**0.01**	**0.01**	**0.001**	**0.001**	**0.001**	**0.001**	**0.01**
6	OP	0.647	0.633	0.638	0.644	0.658	0.667	0.676
	OUM	0.631	0.648	0.636	0.639	0.663	0.684	0.675
	ZP	**0.01**	**0.01**	**0.01**	**0.001**	**0.001**	**0.001**	**0.001**
7	OP	0.644	0.637	0.649	0.651	0.656	0.663	0.666
	OUM	0.696	0.677	0.671	0.708	0.72	0.724	0.727
	ZP	**0.01**	**0.01**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**
8	OP	0.640	0.630	0.623	0.632	0.648	0.658	0.667
	OUM	0.631	0.632	0.747	0.75	0.743	0.736	0.755
	ZP	**0.01**	**0.01**	**0.01**	**0.01**	**0.01**	**0.01**	**0.01**
9	OP	0.635	0.636	0.655	0.666	0.676	0.680	0.684
	OUM	0.673	0.610	0.618	0.643	0.648	0.656	0.660
	ZP	**0.01**	**0.01**	**0.01**	**0.001**	**0.001**	**0.01**	**0.01**
10	OP	0.652	0.636	0.650	0.660	0.668	0.676	0.680
	OUM	0.651	0.657	0.661	0.621	0.629	0.673	0.675
	Ocena	**0.642**	**0.624**	**0.646**	**0.648**	**0.658**	**0.673**	**0.682**

Tabela A.2: Napovedne točnosti na podlagi ocene točnosti za logistično regresijo, izvedeno na komentarjih.

	n-terka	2	3	4	5	6	7	8
	ZP	**0.01**	**0.01**	**1e-06**	**1e-06**	**1e-05**	**0.1**	**0.1**
1	OP	0.628	0.657	0.678	0.639	0.618	0.622	0.612
	OUM	0.575	0.590	0.666	0.606	0.651	0.621	0.636
	ZP	**0.01**	**0.001**	**1e-04**	**1e-06**	**0.1**	**0.1**	**0.1**
2	OP	0.623	0.654	0.656	0.639	0.614	0.610	0.611
	OUM	0.666	0.714	0.690	0.738	0.666	0.666	0.714
	ZP	**1e-05**	**1e-05**	**1e-06**	**1e-06**	**1e-04**	**1e-04**	**1e-05**
3	OP	0.621	0.665	0.680	0.667	0.653	0.652	0.642
	OUM	0.661	0.661	0.693	0.677	0.661	0.661	0.596
	ZP	**0.01**	**0.01**	**1e-04**	**0.001**	**1e-04**	**0.01**	**1e-04**
4	OP	0.630	0.638	0.647	0.630	0.631	0.632	0.642
	OUM	0.735	0.792	0.603	0.641	0.641	0.622	0.547
	ZP	**0.001**	**0.01**	**1e-04**	**1e-04**	**0.01**	**0.1**	**0.1**
5	OP	0.655	0.667	0.662	0.646	0.635	0.636	0.617
	OUM	0.6	0.618	0.618	0.563	0.581	0.581	0.563
	ZP	**0.01**	**0.01**	**0.001**	**1e-04**	**0.1**	**0.001**	**0.01**
6	OP	0.638	0.656	0.648	0.635	0.613	0.612	0.619
	OUM	0.6	0.625	0.625	0.675	0.65	0.65	0.625
	ZP	**0.01**	**0.01**	**0.01**	**1e-05**	**1e-06**	**1e-05**	**1e-05**
7	OP	0.642	0.651	0.643	0.647	0.650	0.656	0.650
	OUM	0.566	0.566	0.566	0.583	0.583	0.566	0.566
	ZP	**0.01**	**0.001**	**1e-06**	**0.001**	**0.001**	**1e-05**	**0.001**
8	OP	0.639	0.679	0.670	0.657	0.660	0.657	0.642
	OUM	0.725	0.627	0.607	0.568	0.568	0.588	0.568
	ZP	**0.01**	**0.001**	**0.1**	**1e-06**	**0.001**	**0.01**	**1e-05**
9	OP	0.640	0.664	0.630	0.636	0.622	0.629	0.622
	OUM	0.588	0.627	0.627	0.666	0.686	0.627	0.705
	ZP	**0.01**	**0.01**	**0.01**	**1e-05**	**1e-04**	**0.01**	**1e-04**
10	OP	0.613	0.674	0.650	0.629	0.633	0.631	0.617
	OUM	0.683	0.65	0.65	0.7	0.616	0.6	0.616
	Ocena	**0.640**	**0.647**	**0.635**	**0.642**	**0.631**	**0.619**	**0.614**

Tabela A.3: Napovedne točnosti na podlagi ocene logloss za logistično regresijo, izvedeno na žanrih.

	n-terka	2	3	4	5	6	7	8
1	ZP	**0.001**	**1e-06**	**1e-10**	**1e-10**	**1e-10**	**1e-10**	**1e-10**
	OP	0.349	0.133	0.141	0.191	0.259	0.329	0.401
	OUM	0.359	0.168	0.173	0.212	0.273	0.333	0.396
2	ZP	**1e-04**	**1e-10**	**1e-10**	**1e-10**	**1e-10**	**1e-10**	**1e-10**
	OP	0.360	0.131	0.143	0.181	0.241	0.323	0.392
	OUM	0.217	0.079	0.087	0.152	0.167	0.261	0.335
3	ZP	**0.001**	**1e-06**	**1e-10**	**1e-10**	**1e-10**	**1e-10**	**1e-10**
	OP	0.348	0.133	0.145	0.201	0.260	0.326	0.395
	OUM	0.417	0.169	0.132	0.171	0.234	0.302	0.375
4	ZP	**0.001**	**1e-10**	**1e-10**	**1e-10**	**1e-10**	**1e-10**	**1e-10**
	OP	0.385	0.154	0.152	0.198	0.262	0.329	0.400
	OUM	0.261	0.065	0.094	0.160	0.215	0.284	0.345
5	ZP	**0.001**	**1e-05**	**1e-10**	**1e-10**	**1e-10**	**1e-10**	**1e-10**
	OP	0.355	0.147	0.144	0.190	0.259	0.326	0.394
	OUM	0.300	0.097	0.128	0.170	0.238	0.300	0.368
6	ZP	**0.001**	**1e-10**	**1e-10**	**1e-10**	**1e-10**	**1e-10**	**1e-10**
	OP	0.335	0.129	0.139	0.180	0.243	0.320	0.387
	OUM	0.439	0.134	0.126	0.172	0.243	0.318	0.392
7	ZP	**1e-04**	**1e-10**	**1e-10**	**1e-10**	**1e-10**	**1e-10**	**1e-10**
	OP	0.320	0.137	0.144	0.191	0.258	0.332	0.399
	OUM	0.278	0.164	0.149	0.170	0.204	0.260	0.335
8	ZP	**0.001**	**1e-10**	**1e-10**	**1e-10**	**1e-10**	**1e-10**	**1e-10**
	OP	0.329	0.133	0.137	0.184	0.240	0.310	0.378
	OUM	0.442	0.150	0.161	0.24	0.297	0.418	0.483
9	ZP	**1e-04**	**1e-10**	**1e-10**	**1e-10**	**1e-10**	**1e-10**	**1e-10**
	OP	0.326	0.136	0.147	0.208	0.267	0.335	0.403
	OUM	0.477	0.127	0.139	0.176	0.230	0.299	0.367
10	ZP	**0.001**	**1e-10**	**1e-10**	**1e-10**	**1e-10**	**1e-10**	**1e-10**
	OP	0.358	0.148	0.140	0.188	0.256	0.331	0.405
	OUM	0.271	0.106	0.106	0.116	0.156	0.207	0.278
	Ocena	**0.347**	**0.126**	**0.130**	**0.174**	**0.226**	**0.299**	**0.368**

Tabela A.4: Napovedne točnosti na podlagi ocene točnosti za logistično regresijo, izvedeno na žanrih.

	n-terka	2	3	4	5	6	7	8
1	ZP	**0.001**	**1e-06**	**1e-04**	**1e-05**	**1e-05**	**1e-05**	**1e-04**
	OP	0.860	0.964	0.976	0.967	0.956	0.946	0.924
	OUM	0.857	0.918	0.959	0.959	0.959	0.938	0.938
2	ZP	**1e-04**	**1e-06**	**0.001**	**1e-05**	**1e-04**	**1e-05**	**1e-05**
	OP	0.854	0.955	0.965	0.970	0.954	0.943	0.930
	OUM	0.939	0.969	1.0	1.0	1.0	0.939	0.939
3	ZP	**1e-04**	**0.01**	**1e-05**	**1e-05**	**1e-05**	**1e-05**	**1e-05**
	OP	0.857	0.955	0.971	0.969	0.955	0.938	0.915
	OUM	0.822	0.933	0.955	0.955	0.955	0.955	0.977
4	ZP	**0.001**	**0.001**	**0.001**	**1e-04**	**1e-06**	**1e-05**	**1e-07**
	OP	0.834	0.946	0.964	0.956	0.943	0.941	0.921
	OUM	0.921	0.973	1.0	0.973	1.0	0.947	0.973
5	ZP	**1e-05**	**1e-05**	**0.01**	**0.001**	**1e-04**	**0.01**	**1e-04**
	OP	0.843	0.950	0.971	0.970	0.967	0.946	0.929
	OUM	0.942	0.971	0.914	0.914	0.914	0.885	0.914
6	ZP	**1e-06**	**1e-04**	**0.001**	**1e-04**	**1e-04**	**0.001**	**0.001**
	OP	0.868	0.953	0.974	0.963	0.950	0.930	0.922
	OUM	0.805	0.972	0.972	1.0	0.972	0.972	0.944
7	ZP	**1e-04**	**1e-04**	**1e-04**	**1e-04**	**1e-06**	**1e-06**	**1e-06**
	OP	0.872	0.948	0.971	0.974	0.953	0.941	0.934
	OUM	0.931	0.954	0.954	0.977	0.954	0.954	0.954
8	ZP	**0.001**	**0.001**	**0.001**	**0.001**	**1e-05**	**1e-05**	**1e-05**
	OP	0.879	0.953	0.968	0.973	0.954	0.935	0.928
	OUM	0.761	0.904	0.952	0.976	0.952	0.928	0.857
9	ZP	**1e-04**	**0.001**	**1e-06**	**1e-05**	**0.1**	**1e-05**	**0.001**
	OP	0.861	0.956	0.969	0.964	0.952	0.932	0.911
	OUM	0.789	0.947	0.921	0.973	0.921	0.921	0.921
10	ZP	**0.001**	**0.001**	**1e-04**	**1e-05**	**1e-04**	**1e-05**	**0.1**
	OP	0.853	0.948	0.975	0.978	0.965	0.936	0.912
	OUM	0.871	0.974	0.974	0.974	0.974	0.974	0.948
	Ocena	**0.864**	**0.952**	**0.960**	**0.970**	**0.960**	**0.942**	**0.937**

Dodatek B

Metoda podpornih vektorjev - rezultati

Tabela B.1: Napovedne točnosti na podlagi ocene logloss za metodo podpornih vektorjev, izvedeno na komentarjih.

	n-terka	2	3	4	5	6	7	8
	ZP	**200**	**200**	**200**	**200**	**200**	**200**	**200**
1	OP	0.635	0.638	0.637	0.649	0.664	0.671	0.676
	OUM	0.680	0.636	0.639	0.646	0.653	0.662	0.663
	ZP	**200**	**200**	**200**	**200**	**200**	**200**	**200**
2	OP	0.653	0.640	0.643	0.647	0.656	0.663	0.668
	OUM	0.606	0.564	0.569	0.586	0.609	0.627	0.634
	ZP	**200**	**200**	**200**	**200**	**200**	**200**	**200**
3	OP	0.661	0.629	0.612	0.624	0.640	0.651	0.657
	OUM	0.600	0.603	0.642	0.636	0.629	0.631	0.639
	ZP	**200**	**200**	**200**	**200**	**200**	**200**	**200**
4	OP	0.658	0.654	0.647	0.649	0.653	0.659	0.664
	OUM	0.602	0.604	0.608	0.613	0.625	0.633	0.640
	ZP	**200**	**200**	**200**	**200**	**200**	**700**	**200**
5	OP	0.655	0.625	0.622	0.635	0.645	0.651	0.655
	OUM	0.649	0.619	0.654	0.653	0.657	0.662	0.671
	ZP	**200**	**200**	**200**	**200**	**200**	**200**	**200**
6	OP	0.647	0.641	0.634	0.641	0.650	0.657	0.663
	OUM	0.626	0.654	0.639	0.644	0.658	0.668	0.671
	ZP	**400**	**200**	**200**	**200**	**200**	**700**	**700**
7	OP	0.657	0.644	0.642	0.642	0.647	0.652	0.656
	OUM	0.684	0.669	0.636	0.664	0.675	0.690	0.692
	ZP	**200**	**200**	**200**	**200**	**200**	**700**	**700**
8	OP	0.649	0.630	0.621	0.630	0.642	0.650	0.657
	OUM	0.640	0.636	0.700	0.733	0.723	0.736	0.745
	ZP	**200**	**200**	**200**	**200**	**200**	**200**	**700**
9	OP	0.643	0.634	0.652	0.659	0.669	0.672	0.673
	OUM	0.677	0.633	0.607	0.626	0.638	0.650	0.656
	ZP	**200**	**200**	**200**	**200**	**200**	**200**	**200**
10	OP	0.654	0.641	0.651	0.653	0.657	0.661	0.665
	OUM	0.640	0.648	0.633	0.630	0.645	0.658	0.6659
	Ocena	**0.641**	**0.627**	**0.633**	**0.644**	**0.652**	**0.662**	**0.668**

Tabela B.2: Napovedne točnosti na podlagi ocene točnosti za metodo podpornih vektorjev, izvedeno na komentarjih.

	n-terka	2	3	4	5	6	7	8
1	ZP	**400**	**200**	**200**	**200**	**200**	**200**	**300**
	OP	0.654	0.658	0.638	0.612	0.61	0.606	0.605
	OUM	0.560	0.606	0.636	0.606	0.590	0.590	0.590
2	ZP	**200**	**200**	**200**	**200**	**200**	**200**	**200**
	OP	0.639	0.637	0.631	0.634	0.614	0.615	0.609
	OUM	0.595	0.690	0.666	0.690	0.666	0.666	0.666
3	ZP	**200**	**300**	**400**	**500**	**600**	**400**	**300**
	OP	0.613	0.658	0.673	0.664	0.637	0.624	0.617
	OUM	0.758	0.709	0.677	0.693	0.677	0.677	0.661
4	ZP	**200**	**200**	**200**	**200**	**500**	**300**	**400**
	OP	0.628	0.646	0.635	0.628	0.619	0.606	0.602
	OUM	0.773	0.679	0.698	0.679	0.660	0.641	0.622
5	ZP	**200**	**300**	**200**	**300**	**600**	**500**	**200**
	OP	0.636	0.661	0.660	0.652	0.634	0.623	0.612
	OUM	0.563	0.618	0.545	0.581	0.618	0.6	0.6
6	ZP	**300**	**300**	**200**	**300**	**200**	**200**	**200**
	OP	0.641	0.651	0.644	0.632	0.624	0.611	0.612
	OUM	0.675	0.6	0.65	0.65	0.675	0.625	0.625
7	ZP	**600**	**200**	**200**	**200**	**200**	**500**	**200**
	OP	0.630	0.645	0.641	0.641	0.634	0.622	0.614
	OUM	0.583	0.6	0.583	0.583	0.616	0.6	0.6
8	ZP	**200**	**200**	**700**	**400**	**300**	**600**	**200**
	OP	0.625	0.654	0.657	0.666	0.649	0.641	0.628
	OUM	0.647	0.647	0.588	0.568	0.588	0.529	0.529
9	ZP	**500**	**200**	**400**	**200**	**300**	**400**	**200**
	OP	0.636	0.647	0.631	0.625	0.609	0.604	0.598
	OUM	0.568	0.686	0.725	0.647	0.627	0.607	0.627
10	ZP	**600**	**200**	**200**	**300**	**300**	**300**	**300**
	OP	0.625	0.652	0.636	0.634	0.631	0.619	0.615
	OUM	0.616	0.6	0.65	0.616	0.616	0.6	0.583
	Ocena	**0.634**	**0.644**	**0.642**	**0.632**	**0.634**	**0.614**	**0.611**

Tabela B.3: Napovedne točnosti na podlagi ocene logloss za metodo podpornih vektorjev, izvedeno na žanrih.

	n-terka	2	3	4	5	6	7	8
	ZP	**200**	**200**	**400**	**300**	**300**	**300**	**300**
1	OP	0.340	0.131	0.092	0.083	0.098	0.142	0.182
	OUM	0.302	0.145	0.105	0.105	0.153	0.219	0.257
	ZP	**200**	**500**	**400**	**300**	**300**	**300**	**300**
2	OP	0.337	0.135	0.093	0.084	0.105	0.148	0.188
	OUM	0.277	0.091	0.028	0.035	0.033	0.062	0.088
	ZP	**200**	**200**	**400**	**300**	**300**	**300**	**300**
3	OP	0.327	0.118	0.096	0.093	0.115	0.158	0.200
	OUM	0.468	0.171	0.064	0.047	0.076	0.099	0.128
	ZP	**200**	**200**	**400**	**300**	**300**	**300**	**300**
4	OP	0.382	0.154	0.108	0.095	0.109	0.147	0.191
	OUM	0.238	0.068	0.033	0.042	0.053	0.094	0.096
	ZP	**200**	**200**	**400**	**300**	**300**	**300**	**300**
5	OP	0.337	0.148	0.093	0.088	0.114	0.158	0.202
	OUM	0.278	0.109	0.093	0.102	0.100	0.136	0.195
	ZP	**200**	**200**	**400**	**300**	**300**	**300**	**300**
6	OP	0.314	0.121	0.086	0.085	0.112	0.158	0.197
	OUM	0.428	0.088	0.069	0.047	0.058	0.089	0.128
	ZP	**200**	**400**	**300**	**300**	**300**	**300**	**300**
7	OP	0.328	0.123	0.081	0.088	0.107	0.149	0.185
	OUM	0.257	0.178	0.129	0.120	0.122	0.130	0.144
	ZP	**200**	**200**	**300**	**300**	**300**	**300**	**300**
8	OP	0.319	0.129	0.097	0.094	0.114	0.150	0.187
	OUM	0.426	0.157	0.082	0.082	0.125	0.215	0.300
	ZP	**200**	**200**	**300**	**300**	**300**	**300**	**300**
9	OP	0.322	0.138	0.101	0.096	0.120	0.165	0.207
	OUM	0.355	0.120	0.114	0.117	0.137	0.175	0.195
	ZP	**200**	**200**	**300**	**300**	**300**	**300**	**300**
10	OP	0.338	0.142	0.083	0.084	0.107	0.149	0.199
	OUM	0.283	0.092	0.080	0.050	0.047	0.067	0.106
	Ocena	**0.332**	**0.122**	**0.080**	**0.075**	**0.091**	**0.129**	**0.164**

Tabela B.4: Napovedne točnosti na podlagi ocene točnosti za metodo podpornih vektorjev, izvedeno na žanrih.

	n-terka	2	3	4	5	6	7	8
	ZP	**200**	**200**	**200**	**200**	**200**	**200**	**200**
1	OP	0.852	0.956	0.969	0.967	0.958	0.947	0.938
	OUM	0.857	0.938	0.959	0.959	0.938	0.938	0.918
	ZP	**400**	**200**	**300**	**200**	**300**	**300**	**200**
2	OP	0.861	0.944	0.972	0.967	0.964	0.955	0.929
	OUM	0.909	0.939	1.0	1.0	1.0	1.0	0.969
	ZP	**700**	**400**	**300**	**200**	**200**	**300**	**300**
3	OP	0.861	0.953	0.974	0.969	0.960	0.946	0.930
	OUM	0.755	0.933	0.955	0.977	0.955	0.955	0.955
	ZP	**200**	**200**	**200**	**300**	**200**	**200**	**200**
4	OP	0.836	0.942	0.952	0.958	0.960	0.950	0.929
	OUM	0.921	0.973	1.0	1.0	1.0	0.973	0.973
	ZP	**400**	**200**	**200**	**200**	**200**	**200**	**200**
5	OP	0.832	0.948	0.969	0.970	0.966	0.956	0.938
	OUM	0.914	0.971	0.971	0.942	0.971	0.942	0.942
	ZP	**200**	**200**	**300**	**300**	**300**	**200**	**200**
6	OP	0.868	0.957	0.974	0.963	0.950	0.943	0.934
	OUM	0.777	0.972	0.944	1.0	0.972	1.0	0.972
	ZP	**600**	**200**	**200**	**200**	**200**	**200**	**300**
7	OP	0.857	0.950	0.968	0.972	0.960	0.951	0.939
	OUM	0.954	0.954	0.954	0.977	0.931	0.931	0.931
	ZP	**500**	**300**	**200**	**200**	**200**	**300**	**200**
8	OP	0.868	0.950	0.970	0.971	0.966	0.954	0.942
	OUM	0.785	0.928	0.976	0.976	0.976	0.928	0.833
	ZP	**400**	**200**	**200**	**200**	**200**	**200**	**200**
9	OP	0.862	0.957	0.971	0.967	0.951	0.948	0.926
	OUM	0.842	0.921	0.921	0.921	0.973	0.947	0.947
	ZP	**200**	**300**	**200**	**200**	**200**	**300**	**200**
10	OP	0.866	0.953	0.980	0.975	0.967	0.941	0.921
	OUM	0.846	0.974	0.974	0.974	0.974	0.974	0.974
	Ocena	**0.856**	**0.951**	**0.966**	**0.973**	**0.969**	**0.959**	**0.942**

Dodatek C

Metoda k najbližjih sosedov - rezultati

Tabela C.1: Napovedne točnosti na podlagi ocene logloss za metodo k najbližjih sosedov, izvedeno na komentarjih.

	n-terka	2	3	4	5	6	7	8
	ZP	**50**	**50**	**8**	**12**	**40**	**20**	**100**
1	OP	0.668	0.672	0.659	0.658	0.669	0.683	0.694
	OUM	0.671	0.666	0.666	0.683	0.688	0.665	0.710
	ZP	**50**	**13**	**30**	**30**	**40**	**40**	**30**
2	OP	0.667	0.661	0.663	0.669	0.670	0.680	0.687
	OUM	0.606	0.596	0.629	0.617	0.649	0.643	0.669
	ZP	**40**	**20**	**30**	**30**	**40**	**100**	**100**
3	OP	0.669	0.663	0.670	0.668	0.670	0.684	0.686
	OUM	0.671	0.603	0.650	0.663	0.629	0.641	0.641
	ZP	**50**	**30**	**20**	**30**	**20**	**30**	**30**
4	OP	0.664	0.663	0.667	0.674	0.667	0.678	0.687
	OUM	0.644	0.673	0.718	0.662	0.651	0.724	0.698
	ZP	**50**	**100**	**8**	**50**	**30**	**12**	**100**
5	OP	0.665	0.672	0.664	0.673	0.672	0.683	0.690
	OUM	0.654	0.658	0.718	0.658	0.658	0.694	0.673
	ZP	**40**	**100**	**20**	**10**	**30**	**40**	**100**
6	OP	0.655	0.667	0.668	0.660	0.668	0.675	0.692
	OUM	0.664	0.652	0.707	0.741	0.700	0.695	0.694
	ZP	**40**	**13**	**13**	**30**	**40**	**20**	**50**
7	OP	0.657	0.664	0.669	0.673	0.671	0.679	0.679
	OUM	0.676	0.684	0.660	0.686	0.727	0.715	0.756
	ZP	**40**	**8**	**10**	**13**	**20**	**30**	**100**
8	OP	0.659	0.645	0.644	0.653	0.655	0.676	0.690
	OUM	0.668	0.660	0.729	0.708	0.690	0.696	0.695
	ZP	**30**	**100**	**12**	**10**	**50**	**30**	**30**
9	OP	0.663	0.667	0.660	0.665	0.673	0.676	0.682
	OUM	0.718	0.700	0.668	0.721	0.700	0.729	0.745
	ZP	**30**	**20**	**20**	**30**	**40**	**20**	**30**
10	OP	0.674	0.664	0.666	0.670	0.679	0.687	0.684
	OUM	0.676	0.672	0.639	0.674	0.660	0.647	0.686
	Ocena	**0.665**	**0.657**	**0.679**	**0.682**	**0.676**	**0.685**	**0.697**

Tabela C.2: Napovedne točnosti na podlagi ocene točnosti za metodo k najbližjih sosedov, izvedeno na komentarjih.

	n-terka	2	3	4	5	6	7	8
	ZP	**8**	**12**	**12**	**10**	**40**	**6**	**12**
1	OP	0.618	0.628	0.631	0.619	0.610	0.616	0.608
	OUM	0.560	0.545	0.545	0.590	0.590	0.560	0.590
	ZP	**6**	**13**	**13**	**10**	**12**	**8**	**40**
2	OP	0.632	0.622	0.613	0.607	0.618	0.605	0.599
	OUM	0.619	0.738	0.714	0.714	0.690	0.690	0.666
	ZP	**10**	**10**	**8**	**20**	**10**	**4**	**4**
3	OP	0.604	0.621	0.613	0.615	0.608	0.620	0.616
	OUM	0.629	0.661	0.709	0.677	0.725	0.677	0.661
	ZP	**4**	**13**	**10**	**4**	**12**	**10**	**12**
4	OP	0.620	0.629	0.624	0.616	0.618	0.605	0.629
	OUM	0.641	0.566	0.622	0.622	0.603	0.641	0.603
	ZP	**20**	**4**	**13**	**30**	**30**	**8**	**4**
5	OP	0.615	0.627	0.618	0.614	0.608	0.614	0.615
	OUM	0.636	0.545	0.654	0.618	0.636	0.618	0.618
	ZP	**6**	**10**	**8**	**8**	**8**	**6**	**6**
6	OP	0.614	0.626	0.623	0.615	0.630	0.628	0.624
	OUM	0.55	0.65	0.6	0.6	0.6	0.625	0.55
	ZP	**20**	**12**	**13**	**20**	**13**	**12**	**8**
7	OP	0.616	0.627	0.627	0.617	0.611	0.616	0.612
	OUM	0.583	0.616	0.633	0.566	0.583	0.616	0.566
	ZP	**40**	**6**	**8**	**8**	**8**	**8**	**8**
8	OP	0.621	0.647	0.653	0.621	0.629	0.650	0.650
	OUM	0.588	0.588	0.549	0.568	0.490	0.549	0.568
	ZP	**20**	**50**	**12**	**4**	**8**	**12**	**10**
9	OP	0.622	0.611	0.626	0.626	0.630	0.628	0.626
	OUM	0.568	0.549	0.588	0.549	0.549	0.568	0.529
	ZP	**10**	**12**	**13**	**8**	**6**	**40**	**40**
10	OP	0.617	0.633	0.615	0.612	0.609	0.609	0.611
	OUM	0.583	0.616	0.65	0.65	0.583	0.583	0.583
	Ocena	**0.596**	**0.608**	**0.627**	**0.616**	**0.605**	**0.613**	**0.594**

Tabela C.3: Napovedne točnosti na podlagi ocene logloss za metodo k najbližjih sosedov, izvedeno na žanrih.

	n-terka	2	3	4	5	6	7	8
	ZP	**10**	**8**	**10**	**10**	**8**	**20**	**20**
1	OP	0.466	0.358	0.390	0.408	0.353	0.482	0.486
	OUM	0.370	0.327	0.364	0.384	0.306	0.443	0.452
	ZP	**10**	**10**	**8**	**6**	**8**	**10**	**12**
2	OP	0.459	0.345	0.334	0.322	0.351	0.371	0.417
	OUM	0.356	0.293	0.230	0.240	0.213	0.260	0.317
	ZP	**13**	**10**	**8**	**12**	**10**	**8**	**12**
3	OP	0.461	0.352	0.331	0.385	0.359	0.333	0.458
	OUM	0.487	0.401	0.368	0.414	0.372	0.373	0.411
	ZP	**8**	**6**	**10**	**10**	**10**	**10**	**8**
4	OP	0.453	0.302	0.371	0.367	0.361	0.379	0.461
	OUM	0.424	0.261	0.240	0.283	0.267	0.296	0.298
	ZP	**12**	**10**	**8**	**10**	**10**	**10**	**8**
5	OP	0.457	0.356	0.329	0.349	0.342	0.348	0.441
	OUM	0.449	0.267	0.303	0.405	0.295	0.382	0.326
	ZP	**6**	**10**	**8**	**10**	**8**	**10**	**12**
6	OP	0.406	0.339	0.340	0.369	0.321	0.366	0.413
	OUM	139.957	0.364	0.369	0.361	0.321	0.383	0.327
	ZP	**20**	**12**	**8**	**10**	**8**	**8**	**8**
7	OP	0.504	0.379	0.360	0.395	0.342	0.379	0.370
	OUM	0.396	0.288	0.269	0.258	0.209	0.254	0.289
	ZP	**8**	**8**	**6**	**6**	**6**	**10**	**12**
8	OP	0.402	0.301	0.278	0.282	0.269	0.351	0.389
	OUM	0.566	0.436	0.458	1.255	1.244	0.520	0.471
	ZP	**8**	**8**	**8**	**6**	**8**	**8**	**8**
9	OP	0.433	0.323	0.350	0.335	0.327	0.347	0.335
	OUM	0.478	0.313	0.355	0.331	0.338	0.348	0.350
	ZP	**10**	**8**	**6**	**8**	**8**	**20**	**10**
10	OP	0.461	0.332	0.295	0.344	0.339	0.474	0.384
	OUM	0.303	0.178	0.127	0.160	0.153	0.292	1.075
	Ocena	**0.523**	**0.313**	**0.309**	**0.410**	**0.372**	**0.356**	**0.432**

Tabela C.4: Napovedne točnosti na podlagi ocene točnosti za metodo k najbližjih sosedov, izvedeno na žanrih.

	n-terka	2	3	4	5	6	7	8
1	ZP	**10**	**4**	**6**	**4**	**4**	**4**	**4**
	OP	0.819	0.891	0.878	0.878	0.887	0.875	0.870
	OUM	0.795	0.918	0.857	0.897	0.897	0.918	0.897
2	ZP	**10**	**6**	**6**	**4**	**4**	**6**	**4**
	OP	0.814	0.907	0.909	0.900	0.895	0.864	0.867
	OUM	0.909	0.909	0.909	0.939	0.909	0.939	0.939
3	ZP	**8**	**6**	**4**	**4**	**4**	**4**	**4**
	OP	0.828	0.923	0.912	0.897	0.903	0.879	0.858
	OUM	0.777	0.888	0.844	0.933	0.888	0.911	0.933
4	ZP	**6**	**8**	**4**	**4**	**4**	**4**	**4**
	OP	0.797	0.900	0.887	0.886	0.888	0.880	0.870
	OUM	0.842	0.947	0.947	0.947	0.947	0.921	0.842
5	ZP	**4**	**6**	**6**	**4**	**4**	**4**	**4**
	OP	0.830	0.909	0.904	0.901	0.918	0.897	0.892
	OUM	0.828	0.942	0.828	0.885	0.828	0.828	0.885
6	ZP	**4**	**4**	**4**	**4**	**6**	**4**	**4**
	OP	0.837	0.937	0.905	0.892	0.890	0.878	0.873
	OUM	0.777	0.944	0.861	0.916	0.944	0.861	0.944
7	ZP	**12**	**4**	**8**	**4**	**6**	**4**	**6**
	OP	0.818	0.908	0.878	0.888	0.887	0.871	0.870
	OUM	0.909	0.931	0.931	0.931	0.886	0.863	0.886
8	ZP	**6**	**4**	**6**	**4**	**4**	**6**	**4**
	OP	0.866	0.933	0.899	0.911	0.897	0.879	0.883
	OUM	0.738	0.833	0.809	0.761	0.785	0.785	0.809
9	ZP	**20**	**4**	**4**	**4**	**4**	**4**	**4**
	OP	0.816	0.906	0.891	0.899	0.890	0.879	0.875
	OUM	0.789	0.921	0.868	0.868	0.815	0.894	0.921
10	ZP	**30**	**4**	**4**	**4**	**4**	**6**	**4**
	OP	0.829	0.928	0.921	0.907	0.880	0.884	0.888
	OUM	0.923	0.948	0.948	0.974	0.948	0.948	0.948
	Ocena	**0.829**	**0.919**	**0.881**	**0.906**	**0.885**	**0.887**	**0.901**

Dodatek D

Metoda naključnih gozdov - rezultati

Tabela D.1: Napovedne točnosti na podlagi ocene logloss za metodo naključnih gozdov, izvedeno na komentarjih.

	n-terka	2	3	4	5	6	7	8
1	ZP	**350**	**250**	**300**	**350**	**100**	**300**	**100**
	OP	0.641	0.634	0.666	0.724	0.775	0.791	0.789
	OUM	0.621	0.609	0.649	0.699	0.756	0.778	0.730
2	ZP	**300**	**200**	**350**	**350**	**300**	**350**	**150**
	OP	0.639	0.633	0.662	0.711	0.759	0.780	0.753
	OUM	0.590	0.555	0.560	0.614	0.636	0.689	0.709
3	ZP	**350**	**250**	**350**	**100**	**100**	**350**	**150**
	OP	0.643	0.622	0.641	0.706	0.759	0.794	0.791
	OUM	0.597	0.587	0.629	0.660	0.654	0.670	0.690
4	ZP	**100**	**300**	**150**	**250**	**350**	**250**	**250**
	OP	0.642	0.635	0.695	0.811	0.894	0.961	0.922
	OUM	0.678	0.677	0.701	0.797	0.889	0.970	0.876
5	ZP	**150**	**350**	**150**	**350**	**100**	**100**	**100**
	OP	0.628	0.614	0.645	0.708	0.747	0.775	0.760
	OUM	0.657	0.656	0.735	0.740	0.795	0.825	0.818
6	ZP	**350**	**250**	**100**	**100**	**100**	**350**	**100**
	OP	0.639	0.626	0.661	0.722	0.769	0.809	0.791
	OUM	0.625	0.640	0.639	0.650	0.688	0.702	0.713
7	ZP	**350**	**150**	**350**	**200**	**200**	**100**	**100**
	OP	0.622	0.617	0.661	0.717	0.760	0.792	0.773
	OUM	0.680	0.696	0.705	0.759	0.801	0.844	0.848
8	ZP	**250**	**350**	**350**	**350**	**300**	**350**	**350**
	OP	0.637	0.628	0.660	0.703	0.746	0.765	0.749
	OUM	0.621	0.595	0.627	0.724	0.808	0.806	0.825
9	ZP	**300**	**350**	**350**	**350**	**300**	**350**	**350**
	OP	0.638	0.632	0.676	0.742	0.785	0.808	0.781
	OUM	0.616	0.566	0.571	0.634	0.648	0.702	0.697
10	ZP	**350**	**350**	**300**	**350**	**100**	**200**	**100**
	OP	0.640	0.626	0.652	0.704	0.740	0.760	0.743
	OUM	0.645	0.640	0.647	0.704	0.757	0.800	0.794
	Ocena	**0.634**	**0.623**	**0.647**	**0.699**	**0.744**	**0.779**	**0.770**

Tabela D.2: Napovedne točnosti na podlagi ocene točnosti za metodo naključnih gozdov, izvedeno na komentarjih.

	n-terka	2	3	4	5	6	7	8
1	ZP	**200**	**100**	**250**	**100**	**150**	**100**	**200**
	OP	0.674	0.676	0.648	0.634	0.623	0.606	0.606
	OUM	0.681	0.651	0.621	0.621	0.606	0.606	0.590
2	ZP	**300**	**250**	**250**	**350**	**150**	**100**	**150**
	OP	0.636	0.639	0.637	0.630	0.605	0.604	0.604
	OUM	0.714	0.738	0.738	0.738	0.690	0.690	0.690
3	ZP	**350**	**300**	**250**	**100**	**200**	**150**	**250**
	OP	0.684	0.680	0.656	0.621	0.618	0.613	0.609
	OUM	0.725	0.661	0.693	0.661	0.661	0.629	0.629
4	ZP	**350**	**350**	**300**	**100**	**200**	**300**	**300**
	OP	0.659	0.654	0.633	0.618	0.610	0.605	0.604
	OUM	0.679	0.679	0.679	0.622	0.584	0.584	0.603
5	ZP	**150**	**350**	**300**	**350**	**150**	**150**	**150**
	OP	0.678	0.681	0.665	0.645	0.629	0.622	0.612
	OUM	0.636	0.581	0.6	0.6	0.618	0.618	0.618
6	ZP	**150**	**250**	**300**	**150**	**100**	**200**	**150**
	OP	0.668	0.680	0.656	0.640	0.627	0.610	0.606
	OUM	0.675	0.65	0.7	0.65	0.625	0.65	0.625
7	ZP	**300**	**150**	**300**	**150**	**350**	**100**	**100**
	OP	0.680	0.682	0.664	0.651	0.630	0.616	0.618
	OUM	0.6	0.666	0.633	0.616	0.616	0.6	0.583
8	ZP	**350**	**250**	**350**	**350**	**150**	**150**	**100**
	OP	0.659	0.689	0.643	0.635	0.616	0.611	0.609
	OUM	0.647	0.666	0.588	0.568	0.568	0.549	0.529
9	ZP	**300**	**350**	**150**	**150**	**300**	**100**	**150**
	OP	0.666	0.672	0.645	0.615	0.607	0.603	0.598
	OUM	0.686	0.686	0.705	0.627	0.647	0.627	0.607
10	ZP	**350**	**350**	**150**	**100**	**100**	**300**	**150**
	OP	0.661	0.693	0.665	0.643	0.619	0.610	0.612
	OUM	0.633	0.616	0.633	0.65	0.633	0.633	0.616
	Ocena	**0.668**	**0.660**	**0.659**	**0.636**	**0.625**	**0.619**	**0.610**

Tabela D.3: Napovedne točnosti na podlagi ocene logloss za metodo naključnih gozdov, izvedeno na žanrih.

	n-terka	2	3	4	5	6	7	8
	ZP	**350**	**250**	**250**	**350**	**200**	**300**	**350**
1	OP	0.505	0.385	0.321	0.319	0.365	0.438	0.508
	OUM	0.504	0.381	0.322	0.362	0.450	0.559	0.624
	ZP	**200**	**350**	**100**	**150**	**300**	**350**	**100**
2	OP	0.506	0.391	0.331	0.339	0.378	0.450	0.496
	OUM	0.457	0.356	0.264	0.243	0.266	0.340	0.386
	ZP	**350**	**200**	**200**	**300**	**150**	**200**	**150**
3	OP	0.491	0.390	0.333	0.334	0.373	0.441	0.488
	OUM	0.523	0.389	0.304	0.266	0.301	0.344	0.407
	ZP	**350**	**100**	**100**	**100**	**350**	**300**	**100**
4	OP	0.510	0.390	0.327	0.326	0.371	0.443	0.494
	OUM	0.464	0.345	0.280	0.270	0.291	0.357	0.386
	ZP	**350**	**350**	**150**	**150**	**350**	**300**	**200**
5	OP	0.496	0.388	0.316	0.324	0.377	0.450	0.524
	OUM	0.498	0.385	0.328	0.320	0.362	0.476	0.595
	ZP	**350**	**350**	**100**	**100**	**200**	**350**	**150**
6	OP	0.498	0.381	0.318	0.329	0.370	0.440	0.497
	OUM	0.486	0.371	0.294	0.293	0.327	0.407	0.496
	ZP	**250**	**100**	**150**	**350**	**100**	**200**	**300**
7	OP	0.507	0.393	0.322	0.340	0.391	0.477	0.541
	OUM	0.476	0.366	0.347	0.362	0.421	0.487	0.605
	ZP	**300**	**100**	**250**	**250**	**350**	**300**	**200**
8	OP	0.490	0.389	0.324	0.322	0.358	0.421	0.471
	OUM	0.541	0.413	0.344	0.333	0.372	0.447	0.498
	ZP	**350**	**250**	**150**	**350**	**100**	**150**	**200**
9	OP	0.502	0.387	0.333	0.343	0.380	0.447	0.499
	OUM	0.504	0.363	0.288	0.255	0.298	0.324	0.365
	ZP	**250**	**150**	**150**	**150**	**200**	**200**	**200**
10	OP	0.496	0.390	0.330	0.342	0.396	0.474	0.544
	OUM	0.474	0.340	0.277	0.291	0.334	0.402	0.446
	Ocena	**0.493**	**0.371**	**0.305**	**0.300**	**0.343**	**0.415**	**0.481**

Tabela D.4: Napovedne točnosti na podlagi ocene točnosti za metodo naključnih gozdov, izvedeno na žanrih.

	n-terka	2	3	4	5	6	7	8
	ZP	**350**	**300**	**250**	**100**	**100**	**250**	**350**
1	OP	0.874	0.955	0.940	0.894	0.842	0.775	0.707
	OUM	0.836	0.938	0.938	0.857	0.795	0.653	0.612
	ZP	**300**	**200**	**250**	**250**	**200**	**350**	**200**
2	OP	0.895	0.951	0.944	0.901	0.851	0.771	0.739
	OUM	0.878	1.0	1.0	0.939	0.878	0.848	0.818
	ZP	**300**	**250**	**300**	**200**	**100**	**250**	**200**
3	OP	0.883	0.951	0.952	0.903	0.855	0.818	0.790
	OUM	0.888	0.933	0.955	0.911	0.866	0.844	0.777
	ZP	**150**	**150**	**250**	**150**	**350**	**300**	**100**
4	OP	0.867	0.944	0.956	0.909	0.858	0.797	0.772
	OUM	0.894	0.973	0.973	0.947	0.894	0.842	0.789
	ZP	**300**	**300**	**100**	**150**	**200**	**100**	**200**
5	OP	0.874	0.949	0.960	0.893	0.842	0.765	0.706
	OUM	0.828	0.885	0.828	0.857	0.8	0.714	0.657
	ZP	**300**	**350**	**100**	**200**	**200**	**250**	**250**
6	OP	0.881	0.949	0.944	0.893	0.843	0.808	0.751
	OUM	0.888	0.972	0.972	0.944	0.833	0.805	0.722
	ZP	**200**	**250**	**150**	**300**	**200**	**250**	**150**
7	OP	0.877	0.940	0.936	0.865	0.813	0.741	0.714
	OUM	0.863	0.909	0.886	0.886	0.840	0.772	0.636
	ZP	**150**	**300**	**150**	**250**	**250**	**300**	**100**
8	OP	0.895	0.950	0.944	0.908	0.878	0.819	0.782
	OUM	0.809	0.952	0.952	0.928	0.833	0.809	0.738
	ZP	**300**	**350**	**200**	**350**	**200**	**350**	**200**
9	OP	0.876	0.949	0.948	0.902	0.860	0.807	0.774
	OUM	0.921	0.947	0.973	0.947	0.947	0.947	0.789
	ZP	**200**	**200**	**150**	**100**	**300**	**100**	**200**
10	OP	0.881	0.937	0.925	0.870	0.815	0.750	0.699
	OUM	0.794	0.948	0.948	0.923	0.897	0.820	0.717
	Ocena	**0.861**	**0.946**	**0.943**	**0.914**	**0.859**	**0.806**	**0.726**

Dodatek E

Skladanje - rezultati

Tabela E.1: Napovedne točnosti na podlagi ocene logloss za metodo skladanja, izvedeno na komentarjih.

	n-terka	2	3	4	5	6	7	8
1	ZP	**0.1**	**0.1**	**0.1**	**0.1**	**0.1**	**0.1**	**0.1**
	OP	0.646	0.593	0.614	0.634	0.644	0.660	0.666
	OUM	0.646	0.665	0.667	0.651	0.644	0.661	0.655
2	ZP	**0.01**	**0.1**	**0.1**	**0.1**	**0.1**	**0.1**	**0.1**
	OP	0.640	0.655	0.646	0.650	0.652	0.664	0.665
	OUM	0.658	0.631	0.662	0.671	0.690	0.697	0.689
3	ZP	**0.01**	**0.1**	**0.1**	**0.1**	**0.1**	**0.01**	**0.01**
	OP	0.664	0.662	0.659	0.668	0.670	0.677	0.678
	OUM	0.605	0.592	0.577	0.587	0.594	0.607	0.622
4	ZP	**0.01**	**0.1**	**0.1**	**0.1**	**0.1**	**0.1**	**0.01**
	OP	0.644	0.646	0.638	0.651	0.662	0.665	0.662
	OUM	0.688	0.656	0.645	0.667	0.687	0.697	0.707
5	ZP	**0.01**	**0.1**	**0.1**	**0.01**	**0.1**	**0.1**	**0.01**
	OP	0.638	0.633	0.645	0.646	0.655	0.659	0.661
	OUM	0.687	0.680	0.654	0.687	0.674	0.674	0.678
	Ocena	**0.657**	**0.645**	**0.641**	**0.653**	**0.658**	**0.668**	**0.671**

Tabela E.2: Napovedne točnosti na podlagi ocene točnosti za metodo skladanja, izvedeno na komentarjih.

	n-terka	2	3	4	5	6	7	8
1	ZP	**0.01**	**0.1**	**0.01**	**0.001**	**0.01**	**0.1**	**0.1**
	OP	0.672	0.665	0.636	0.581	0.640	0.614	0.615
	OUM	0.582	0.591	0.591	0.6	0.678	0.634	0.626
2	ZP	**0.1**	**0.01**	**0.001**	**0.01**	**0.1**	**0.001**	**0.01**
	OP	0.644	0.653	0.617	0.620	0.634	0.629	0.618
	OUM	0.58	0.64	0.59	0.59	0.58	0.58	0.59
3	ZP	**0.01**	**0.001**	**0.01**	**0.1**	**0.01**	**0.1**	**0.1**
	OP	0.601	0.602	0.603	0.616	0.574	0.546	0.578
	OUM	0.71	0.71	0.7	0.68	0.7	0.66	0.66
4	ZP	**0.001**	**0.01**	**0.001**	**0.1**	**0.001**	**0.1**	**0.01**
	OP	0.646	0.642	0.657	0.642	0.621	0.620	0.618
	OUM	0.573	0.598	0.590	0.598	0.540	0.540	0.524
5	ZP	**0.01**	**0.1**	**0.1**	**0.1**	**0.001**	**0.1**	**0.01**
	OP	0.661	0.615	0.584	0.635	0.641	0.615	0.613
	OUM	0.592	0.650	0.631	0.563	0.611	0.601	0.601
	Ocena	**0.608**	**0.638**	**0.621**	**0.606**	**0.622**	**0.604**	**0.601**

Tabela E.3: Napovedne točnosti na podlagi ocene logloss za metodo skladanja, izvedeno na žanrih.

	n-terka	2	3	4	5	6	7	8
1	ZP	**0.1**	**0.01**	**0.01**	**0.01**	**0.01**	**0.01**	**0.01**
	OP	0.476	0.191	0.132	0.141	0.150	0.167	0.185
	OUM	0.388	0.268	0.246	0.242	0.295	0.225	0.274
2	ZP	**0.01**	**0.01**	**0.1**	**0.1**	**0.01**	**0.01**	**0.01**
	OP	0.457	0.234	0.189	0.170	0.215	0.237	0.264
	OUM	0.370	0.103	0.128	0.124	0.149	0.131	0.158
3	ZP	**0.1**	**0.1**	**0.01**	**0.1**	**0.01**	**0.01**	**0.01**
	OP	0.412	0.209	0.109	0.138	0.137	0.179	0.218
	OUM	0.452	0.203	0.161	0.172	0.206	0.220	0.246
4	ZP	**0.1**	**0.01**	**0.01**	**0.1**	**0.01**	**0.01**	**0.01**
	OP	0.488	0.219	0.157	0.132	0.240	0.269	0.276
	OUM	0.423	0.099	0.148	0.133	0.203	0.242	0.235
5	ZP	**0.01**	**0.1**	**0.01**	**0.01**	**0.01**	**0.01**	**0.01**
	OP	0.462	0.213	0.161	0.123	0.161	0.184	0.223
	OUM	0.401	0.202	0.104	0.148	0.163	0.158	0.209
	Ocena	**0.407**	**0.175**	**0.158**	**0.165**	**0.204**	**0.196**	**0.225**

Tabela E.4: Napovedne točnosti na podlagi ocene točnosti za metodo skladanja, izvedeno na žanrih.

	n-terka	2	3	4	5	6	7	8
1	ZP	**0.01**	**0.001**	**0.01**	**0.01**	**0.1**	**0.1**	**0.1**
	OP	0.820	0.960	0.970	0.974	0.962	0.945	0.940
	OUM	0.852	0.931	0.931	0.931	0.943	0.954	0.897
2	ZP	**0.001**	**0.001**	**0.001**	**0.1**	**0.1**	**0.1**	**0.1**
	OP	0.806	0.932	0.959	0.958	0.941	0.926	0.912
	OUM	0.861	0.972	0.986	0.986	0.986	0.958	0.944
3	ZP	**0.01**	**0.1**	**0.01**	**0.01**	**0.1**	**0.1**	**0.001**
	OP	0.858	0.945	0.978	0.977	0.969	0.951	0.948
	OUM	0.828	0.947	0.973	0.960	0.960	0.947	0.921
4	ZP	**0.01**	**0.1**	**1e-04**	**0.001**	**0.1**	**0.01**	**0.001**
	OP	0.828	0.939	0.954	0.946	0.938	0.922	0.905
	OUM	0.860	0.989	0.967	0.978	0.967	0.935	0.892
5	ZP	**0.01**	**0.01**	**0.01**	**0.01**	**0.01**	**0.1**	**0.01**
	OP	0.820	0.932	0.958	0.969	0.961	0.948	0.928
	OUM	0.9	0.928	0.985	0.971	0.942	0.957	0.942
	Ocena	**0.861**	**0.954**	**0.969**	**0.966**	**0.960**	**0.951**	**0.920**